Nucleotide Sequence of *Escherichia coli* Pathogenicity Islands

Cross-Reference to Related Applications

[0001] This application is a divisional of, and claims benefit under 35 U.S.C. § 120 to copending U.S. Patent Application No. 08/976,259, filed November 21, 1997, which in turn claims benefit under 35 U.S.C. § 119(e) to U.S. Provisional Application Nos. 60/061,953, filed on October 14, 1997, and 60/031,626, filed on November 22, 1996. Claimed priority documents are hereby incorporated by reference in its entirety.

Statements as to Rights to Inventions Made Under Federally-sponsored Research and Development

[0002] This invention was made with United States government support awarded by the following agencies:

NIH Grant # AI20323; AI25547.

The United States has certain rights to this invention.

Background of the Invention

Field of the Invention

[0003] The present invention relates to novel genes located in two chromosomal regions within *E. coli* that are associated with virulence. These chromosomal regions are known as pathogenicity islands (PAIs).

Related Background Art

[0004] *Escherichia coli* (*E. coli*) is a normal inhabitant of the intestine of humans and various animals. Pathogenic *E. coli* strains are able to cause infections of the intestine (intestinal *E. coli* strains) and of other organs such as the urinary tract (uropathogenic *E. coli*) or the brain (extraintestinal *E. coli*). Intestinal pathogenic *E. coli* are a well established and leading cause of severe infantile diarrhea in the developing world. Additionally, cases of newborn meningitis and sepsis have been attributed to *E. coli* pathogens.

[0005] In contrast to non-pathogenic isolates, pathogenic *E. coli* produce pathogenicity factors which contribute to the ability of strains to cause infectious diseases (Mühldorfer, I. and Hacker, J., *Microb. Pathogen.* 16:171-181 1994). Adhesions facilitate binding of

pathogenic bacteria to host tissues. Pathogenic *E. coli* strains also express toxins including haemolysins, which are involved in the destruction of host cells, and surface structures such as O-antigens, capsules or membrane proteins, which protect the bacteria from the action of phagocytes or the complement system (Ritter, *et al.*, *Mol. Microbiol.* 17:109-212 1995).

[0006] The genes coding for pathogenicity factors of intestinal *E. coli* are located on large plasmids, phage genomes or on the chromosome. In contrast to intestinal *E. coli*, pathogenicity determinants of uropathogenic and other extraintestinal *E. coli* are, in most cases, located on the chromosome. *Id.*

[0007] Large chromosomal regions in pathogenic bacteria that encode adjacently located virulence genes have been termed *pathogenicity islands* ("PAIs"). PAIs are indicative of large fragments of DNA which comprise a group of virulence genes behaving as a distinct molecular and functional unit much like an island within the bacterial chromosome. For example, intact PAIs appear to transfer between organisms and confer complex virulence properties to the recipient bacteria.

[0008] Chromosomal PAIs in bacterial cells have been described in increasing detail over recent years. For example, J. Hacker and co-workers described two large, unstable regions in the chromosome of uropathogenic *Escherichia coli* strain 536 as PAI-I and PAI-II (Hacker J., *et al.*, *Microbiol. Pathog.* 8:213-25 1990). Hacker found that PAI-I and PAI-II containing virulence regions can be lost by spontaneous deletion due to recombination events. Both of these PAIs were found to encode multiple virulence genes, and their loss resulted in reduced hemolytic activity, serum resistance, mannose-resistant hemagglutination, uroepithelial cell binding, and mouse virulence of the *E. coli.* (Knapp, S *et al.*, *J. Bacteriol.* 168:22-30 1986). Therefore, pathogenicity islands are characterized by their ability to confer complex virulence phenotypes to bacterial cells.

[0009] In addition to *E. coli,* specific deletion of large virulence regions has been observed in other bacteria such as *Yersinia pestis*. For example, Fetherston and co-workers found that a 102-kb region of the *Y. pestis* chromosome lost by spontaneous deletion resulted in the loss of many *Y. pestis* virulence phenotypes. (Fetherston, J.D. and Perry, R.D., *Mol. Microbiol.* 13:697-708 1994, Fetherston, *et al.*, *Mol. Microbiol.* 6:2693-704 1992). In this instance, the deletion appeared to be due to recombination within 2.2-kb repetitive elements at both ends of the 102-kb region.

[0010] It is possible that deletion of PAIs may benefit the organism by modulating bacterial virulence or genome size during infection. PAIs may also represent foreign DNA segments that were acquired during bacterial evolution that conferred important pathogenic properties to the bacteria. Observed flanking repeats, as observed in *Y. pestis*

for example, may suggest a common mechanism by which these virulence genes were integrated into the bacterial chromosomes.

[0011] Integration of the virulence genes into bacterial chromosomes was further elucidated by the discovery and characterization of a locus of enterocyte effacement (the LEE locus) in enteropathogenic *E. coli* (McDaniel, *et al.*, *Proc. Natl Acad. Sci.* (USA) 92:1664-8 1995). The LEE locus comprises 35-kb and encodes many genes required for these bacteria to "invade" and degrade the apical structure of enerocytes causing diarrhea. Although the LEE and PAI-I loci encode different virulence genes, these elements are located at the exact same site in the *E. coli* genome and contain the same DNA sequence within their right-hand ends, thus suggesting a common mechanism for their insertion.

[0012] Besides being found in enteropathogenic *E. coli*, the LEE element is also present in rabbit diarrheal *E. coli*, *Hafnia alvei*, and *Citrobacter freundii* biotype 4280, all of which induce attaching and effacing lesions on the apical face of enterocytes. The LEE locus appears to be inserted in the bacterial chromosome as a discrete molecular and functional virulence unit in the same fashion as PAI-I, PAI-II, and *Yersinia* PAI.

[0013] Along these same lines, a 40-kb *Salmonella typhimurium* PAI was characterized on the bacterial chromosome which encodes genes required for *Salmonella* entry into nonphagocytic epithelial cells of the intestine (Mills, D.M., *et al.*, *Mol. Microbiol.* 15:749-59 1995). Like the LEE element, this PAI confers to *Salmonella* the ability to invade intestinal cells, and hence may likewise be characterized as an "invasion" PAI.

[0014] The pathogenicity islands described above all possess the common feature of conferring complex virulence properties to the recipient bacteria. However, they may be separated into two types by their respective contributions to virulence. PAI-I, PAI-II, and the *Y. pestis* PAI confer multiple virulence phenotypes, while the LEE and the *S. typhimurium* "invasion" PAI encode many genes specifying a single, complex virulence process.

[0015] It is advantageous to characterize closely-related bacteria that contain or do not contain the PAI by the isolation of a discrete molecular and functional unit on the bacterial chromosome. Since the presence versus the absence of essential virulence genes can often distinguish closely-related virulent versus avirulent bacterial strains or species, experiments have been conducted to identify virulence loci and potential PAIs by isolating DNA sequences that are unique to virulent bacteria (Bloch, C.A., *et al.*, *J Bacteriol.* 176:7121-5 1994, Groisman, E.A., *EMBO J.* 12:3779-87 1993).

[0016] At least two PAIs are present in *E. coli* J96. These PAIs, PAI IV and PAI V are linked to tRNA loci but at sites different from those occupied by other known *E. coli* PAIs. Swenson *et al., Infect. and Immun. 64*:3736-3743 (1996).

[0017] The era of true comparative genomics has been ushered in by high through-put genomic sequencing and analysis. The first two complete bacterial genome sequences, those of *Haemophilus influenzae* and *Mycoplasma genitalium* were recently described (Fleischmann, R.D., *et al., Science 269*:496 (1995); Fraser, C.M., *et al., Science 270*:397 (1995)). Large scale DNA sequencing efforts also have produced an extensive collection of sequence data from eukaryotes, including *Homo sapiens* (Adams, M.D., *et al., Nature* 377:3 (1995)) and *Saccharomyces cerevisiae* (Levy, J., *Yeast 10*:1689 (1994)).

[0018] The need continues to exist for the application of high through-put sequencing and analysis to study genomes and subgenomes of infectious organisms. Further, a need exists for genetic markers that can be employed to distinguish closely-related virulent and avirulent strains of a given bacteria.

Summary of the Invention

[0019] The present invention is based on the high through-put, random sequencing of cosmid clones covering two pathogenic islands (PAIs) of uropathogenic *Escherichia coli* strain J96 (O4:K6; *E. coli* J96). PAIs are large fragments of DNA which comprise pathogenicity determinants. PAI IV is located approximately at 64 min (near*pheV*) on the *E. coli* chromosome and is greater than 170 kilobases in size. PAI V is located at approximately 94 min (*atphe*R) on the *E. coli* chromosome and is approximately 106 kb in size. These PAIs differ in location to the PAIs described by Hacker and colleagues for uropathogenic strain 536 (PAI I, 82 minutes {*sel*C} and PAI II, 97 minutes {*leu*X}).

[0020] The location of the PAIs relative to one another and the cosmid clones covering the J96 PAIs is shown in Figure 1. The present invention relates to the nucleotide sequences of 142 fragments of DNA (contigs) covering the PAI IV and PAI V regions of the *E. coli* J96 chromosome. The nucleotide sequences shown in SEQ ID NOs: 1 through 142 were obtained by shotgun sequencing eleven *E. coli* J96 subclones, which were deposited in two pools on September 23, 1996 at the American Type Culture Collection, 12301 Park Lawn Drive, Rockville, Maryland 20852, and given accession numbers 97726 (includes 7 cosmid clones covering PAI (IV) and 97727 (includes 4 cosmid clones covering PAI V). The deposited sets or "pools" of clones are more fully described in Example 1. In addition, *E. coli* strain J96 was also deposited at the American Type Culture Collection on September 23, 1996, and given accession number 98176.

[0021] Three hundred fifty-one open reading frames have been thus far identified in the 142 contigs described by SEQ ID NOs: 1 through 142. Thus, the present invention is directed to isolated nucleic acid molecules comprising open reading frames (ORFs) encoding *E. coli* proteins that are located in two pathogenic island regions of the chromosome of uropathogenic *E. coli* J96.

[0022] The present invention also relates to variants of the nucleic acid molecules of the present invention, which encode portions, analogs or derivatives of *E. coli* J96 PAI proteins. Further embodiments include isolated nucleic acid molecules comprising a polynucleotide having a nucleotide sequence at least 90% identical, and more preferably at least 95%, 96%, 97%, 98% or 99% identical, to the nucleotide sequence of an *E. coli* J96 PAI ORF described herein.

[0023] The present invention also relates to recombinant vectors, which include the isolated nucleic acid molecules of the present invention, host cells containing the recombinant vectors, as well as methods for making such vectors and host cells for *E. coli* J96 PAI protein production by recombinant techniques.

[0024] The invention further provides isolated polypeptides encoded by the *E. coli* J96 PAI ORFs. It will be recognized that some amino acid sequences of the polypeptides described herein can be varied without significant effect on the structure or function of the protein. If such differences in sequence are contemplated, it should be remembered that there will be critical areas on the protein which determine activity. In general, it is possible to replace residues which form the tertiary structure, provided that residues performing a similar function are used. In other instances, the type of residue may be completely unimportant if the alteration occurs at a non-critical region of the protein.

[0025] In another aspect, the invention provides a peptide or polypeptide comprising an epitope-bearing portion of a polypeptide of the invention. The epitope-bearing portion is an immunogenic or antigenic epitope useful for raising antibodies.

[0026] The invention further provides a vaccine comprising one or more *E. coli* J96 PAI antigens together with a pharmaceutically acceptable diluent, carrier, or excipient, wherein the one or more antigens are present in an amount effective to elicit protective antibodies in an animal to pathogenic *E. coli*, such as strain J96.

[0027] The invention also provides a method of eliciting a protective immune response in an animal comprising administering to the animal the above-described vaccine.

[0028] The invention further provides a method for identifying pathogenic *E. coli* in an animal comprising analyzing tissue or body fluid from the animal for one or more of:

(a) polynucleic acids encoding an open reading frame listed in Tables 1-4;

(b) polypeptides encoded for by an open reading frame listed in Tables 1-4; or

(c) antibodies specific to polypeptides encoded for by an open reading frame listed in Tables 1-4.

[0029] The invention further provides a nucleic acid probe for the detection of the presence of one or more *E. coli* PAI nucleic acids (nucleic acids encoding one or more ORFs as listed in Tables 1-4) in a sample from an individual comprising one or more nucleic acid molecules sufficient to specifically detect under stringent hybridization conditions the presence of the above-described molecule in the sample.

[0030] The invention also provides a method of detecting *E. coli* PAI nucleic acids in a sample comprising:

a) contacting the sample with the above-described nucleic acid probe, under conditions such that hybridization occurs, and

b) detecting the presence of the probe bound to an *E. coli* PAI nucleic acid.

[0031] The invention further provides a kit for detecting the presence of one or more *E. coli* PAI nucleic acids in a sample comprising at least one container means having disposed therein the above-described nucleic acid probe.

[0032] The invention also provides a diagnostic kit for detecting the presence of pathogenic *E. coli* in a sample comprising at least one container means having disposed therein one or more of the above-described antibodies.

[0033] The invention also provides a diagnostic kit for detecting the presence of antibodies to pathogenic *E. coli* in a sample comprising at least one container means having disposed therein one or more of the above-described antigens.

Brief Description of the Figures

[0034] **Figure 1** is a schematic diagram of cosmid clones derived from *E. coli* J96 pathogenicity island and map positions of known *E. coli* PAIs (not drawn to scale). The gray bar represents the *E. coli* K-12 chromosome with minute demarcations of PAI junction points located above the bar. *E. coli* J96 overlapping cosmid clones are represented by hatched bars (overlap not drawn to scale) with positions of *hly*, *pap*, and *prs* operons indicated above bar. The PAIs and estimated sizes are shown above and below the K-12 chromosome map.

[0035] **Figure 2** is a block diagram of a computer system 102 that can be used to implement the computer-based systems of present invention.

Detailed Description of the Invention

[0036] The present invention is based on high through-put, random sequencing of a uropathogenic strain of *Escherichia coli*. The DNA sequences of contiguous DNA

fragments covering the pathogenicity islands, PAI IV (also referred to as $PAI_{J96(pheV)}$) and PAI V (also referred to as $PAI_{J96(pheU)}$) from the chromosome of the *E. coli* uropathogenic strain, J96 (04:K6) were determined. The sequences were used for DNA and protein sequence similarity searches of the database.

[0037] The primary nucleotide sequences generated by shotgun sequencing cosmid clones of the PAI IV and PAI V regions of the *E. coli* chromosome are provided in SEQ ID NOs:1 through 142. These sequences represent contiguous fragments of the PAI DNA. As used herein, the"primary sequence" refers to the nucleotide sequence represented by the IUPAC nomenclature system. The present invention provides the nucleotide sequences of SEQ ID NOs:1 through 142, or representative fragments thereof, in a form that can be readily used, analyzed, and interpreted by a skilled artisan. Within these 142 sequences, there have been thus far identified 351 open reading frames (ORFs) that are described in greater detail below.

[0038] As used herein, a "representative fragment" refers to *E. coli* J96 PAI protein-encoding regions (also referred to herein as open reading frames or ORFs), expression modulating fragments, and fragments that can be used to diagnose the presence of *E. coli* in a sample. A non-limiting identification of such representative fragments is provided in Tables 1 through 6. As described in detail below, representative fragments of the present invention further include nucleic acid molecules having a nucleotide sequence at least 95% identical, preferably at least 96%, 97%, 98%, or 99% identical, to an ORF identified in Tables 1 through 6.

[0039] As indicated above, the nucleotide sequence information provided in SEQ ID NOs:1 through 142 was obtained by sequencing cosmid clones covering the PAIs located on the chromosome of *E. coli* J96 using a megabase shotgun sequencing method. The sequences provided in SEQ ID NOs:1 through 142 are highly accurate, although not necessarily a 100% perfect, representation of the nucleotide sequences of contiguous stretches of DNA (contigs) which include the ORFs located on the two pathogenicity islands of *E. coli* J96. As discussed in detail below, using the information provided in SEQ ID NOs:1 through 142 and in Tables 1 through 6 together with routine cloning and sequencing methods, one of ordinary skill in the art would be able to clone and sequence all "representative fragments" of interest including open reading frames (ORFs) encoding a large variety of *E. coli* J96 PAI proteins. In rare instances, this may reveal a nucleotide sequence error present in the nucleotide sequences disclosed in SEQ ID NOs: 1 through 142. Thus, once the present invention is made available (i.e., once the information in SEQ ID NOs: 1 through 142 and in Tables 1 through 6 have been made available), resolving a rare sequencing error would be well within the skill of the art. Nucleotide sequence

editing software is publicly available. For example, Applied Biosystem's (AB) AutoAssembler can be used as an aid during visual inspection of nucleotide sequences.

[0040] Even if all of the rare sequencing errors were corrected, it is predicted that the resulting nucleotide sequences would still be at least about 99.9% identical to the reference nucleotide sequences in SEQ ID NOs: 1 through 142. Thus, the present invention further provides nucleotide sequences that are at least 99.9% identical to the nucleotide sequence of SEQ ID NOs: 1 through 142 in a form which can be readily used, analyzed and interpreted by the skilled artisan. Methods for determining whether a nucleotide sequence is at least 99.9% identical to a reference nucleotide sequence of the present invention are described below.

Nucleic Acid Molecules

[0041] The present invention is directed to isolated nucleic acid fragments of the PAIs of *E. coli* J96. Such fragments include, but are not limited to, nucleic acid molecules encoding polypeptides (hereinafter open reading frames (ORFs)), nucleic acid molecules that modulate the expression of an operably linked ORF (hereinafter expression modulating fragments (EMFs)), and nucleic acid molecules that can be used to diagnose the presence of *E. coli* in a sample (hereinafter diagnostic fragments (DFs)).

[0042] By isolated nucleic acid molecule(s) is intended a nucleic acid molecule, DNA or RNA, that has been removed from its native environment. For example, recombinant DNA molecules contained in a vector are considered isolated for the purposes of the present invention. Further examples of isolated DNA molecules include recombinant DNA molecules maintained in heterologous host cells, purified (partially or substantially) DNA molecules in solution, and nucleic acid molecules produced synthetically. Isolated RNA molecules include *in vitro* RNA transcripts of the DNA molecules of the present invention.

[0043] In one embodiment, *E. coli* J96 PAI DNA can be mechanically sheared to produce fragments about 15-20 kb in length, which can be used to generate an *E. coli* J96 PAI DNA library by insertion into lambda clones as described in Example 1 below. Primers flanking an ORF described in Tables 1 through 6 can then be generated using the nucleotide sequence information provided in SEQ ID NOs: 1 through 142. The polymerase chain reaction (PCR) is then used to amplify and isolate the ORF from the lambda DNA library. PCR cloning is well known in the art. Thus, given SEQ ID NOs: 1 through 142, and Tables 1 through 6, it would be routine to isolate any ORF or other representative fragment of the *E. coli* J96 PAI subgenomes. Isolated nucleic acid

molecules of the present invention include, but are not limited to, single stranded and double stranded DNA, and single stranded RNA, and complements thereof.

[0044] Tables 1 through 6 herein describe ORFs in the *E. coli* J96 PAI cosmid clone library.

[0045] Tables 1 and 3 list, for PAI IV and PAI V, respectively, a number of ORFs that putatively encode a recited protein based on homology matching with protein sequences from an organism listed in the Table. Tables 1 and 3 indicate the location of ORFs (i.e., the position) by reference to its position within the one of the 142 *E. coli* J96 contigs described in SEQ ID NOs: 1 through 142. Column 1 of Tables 1 and 3 provides the Sequence ID Number (SEQ ID NO) of the contig in which a particular open reading frame is located. Column 2 numerically identifies a particular ORF on a particular contig (SEQ ID NO) since many contigs comprise a plurality of ORFs. Columns 3 and 4 indicate an ORF s position in the nucleotide sequence (contig) provided in SEQ ID NOs: 1 through 142 by referring to start and stop positions in the contig sequence. One of ordinary skill in the art will appreciate that the ORFs may be oriented in opposite directions in the *E. coli* chromosome. This is reflected in columns 3 and 4. Column 5 provides a database accession number to a homologous protein identified by a similarity search of public sequence databases (*see*, infra). Column 6 describes the matching protein sequence and the source organism is identified in brackets. Column 7 of Tables 1 and 3 indicates the percent identity of the protein sequence encoded by an ORF to the corresponding protein sequence from the organism appearing in parentheses in the sixth column. Column 8 of Tables 1 and 3 indicates the percent similarity of the protein sequence encoded by an ORF to the corresponding protein sequence from the organism appearing in parentheses in the sixth column. The concepts of percent identity and percent similarity of two polypeptide sequences are well understood in the art and are described in more detail below. Identified genes can frequently be assigned a putative cellular role category adapted from Riley (*see*, Riley, M., *Microbiol. Rev.* 57:862 (1993)). Column 9 of Tables 1 and 3 provides the nucleotide length of the open reading frame.

[0046] Tables 2 and 4, below, provide ORFs of *E. coli* J96 PAI IV and PAI V, respectively, that did not elicit a homology match with a known sequence from either *E. coli* or another organism. As above, the first column in Tables 2 and 4 provides the contig in which the ORF is located and the second column numerically identifies a particular ORF in a particular contig. Columns 3 and 4 identify an ORF s position in one of SEQ ID NOs: 1 through 142 by reference to start and stop nucleotides.

[0047] Tables 5 and 6, below, provide the *E. coli* J96 PAI IV ORFs and PAI V ORFs, respectively, identified by the present inventors that provided a significant match to a

previously published *E. coli* protein. The columns correspond to the columns appearing in Tables 1 and 3.

[0048] Further details concerning the algorithms and criteria used for homology searches are provided in the Examples below. A skilled artisan can readily identify ORFs in the *Escherichia coli* J96 cosmid library other than those listed in Tables 1 through 6, such as ORFs that are overlapping or encoded by the opposite strand of an identified ORF in addition to those ascertainable using the computer-based systems of the present invention.

[0049] Isolated nucleic acid molecules of the present invention include DNA molecules having a nucleotide sequence substantially different than the nucleotide sequence of an ORF described in Tables 1 through 4, but which, due to the degeneracy of the genetic code, still encode a *E. coli* J96 PAI protein. The genetic code is well known in the art. Thus, it would be routine to generate such degenerate variants.

[0050] The present invention further relates to variants of the nucleic acid molecules of the present invention, which encode portions, analogs or derivatives of an *E. coli* protein encoded by an ORF described in Table 1 through 4. Non-naturally occurring variants may be produced using art-known mutagenesis techniques and include those produced by nucleotide substitutions, deletions or additions. The substitutions, deletions or additions may involve one or more nucleotides. The variants may be altered in coding regions, non-coding regions, or both. Alterations in the coding regions may produce conservative or non-conservative amino acid substitutions, deletions or additions. Especially preferred among these are silent substitutions, additions and deletions, which do not alter the properties and activities of the *E. coli* protein or portions thereof. Also especially preferred in this regard are conservative substitutions.

[0051] Further embodiments of the invention include isolated nucleic acid molecules comprising a polynucleotide having a nucleotide sequence at least 90% identical, and more preferably at least 95%, 96%, 97%, 98% or 99% identical, to the nucleotide sequence of an ORF described in Tables 1 through 6, preferably 1 through 4. By a polynucleotide having a nucleotide sequence at least, for example, 95% identical to the reference *E. coli* ORF nucleotide sequence is intended that the nucleotide sequence of the polynucleotide is identical to the reference sequence except that the polynucleotide sequence may include up to five point mutations per each 100 nucleotides of the ORF sequence. In other words, to obtain a polynucleotide having a nucleotide sequence at least 95% identical to a reference ORF nucleotide sequence, up to 5% of the nucleotides in the reference sequence may be deleted or substituted with another nucleotide, or a number of nucleotides up to 5% of the total nucleotides in the reference sequence may be inserted

into the reference sequence. These mutations of the reference sequence may occur at the 5' or 3' terminal positions of the reference nucleotide sequence or anywhere between those terminal positions, interspersed either individually among nucleotides in the reference sequence or in one or more contiguous groups within the reference sequence.

[0052] As a practical matter, whether any particular nucleic acid molecule is at least 90%, 95%, 96%, 97%, 98% or 99% identical to the nucleotide sequence of an *E. coli* J96 PAI ORF can be determined conventionally using known computer programs such as the Bestfit program (Wisconsin Sequence Analysis Package, Version 8 for Unix, Genetics Computer Group, University Research Park, 575 Science Drive, Madison, WI 53711). Bestfit uses the local homology algorithm of Smith and Waterman, *Advances in Applied Mathematics 2*: 482-489 (1981), to find the best segment of homology between two sequences. When using Bestfit or any other sequence alignment program to determine whether a particular sequence is, for instance, 95% identical to a reference sequence according to the present invention, the parameters are set, of course, such that the percentage of identity is calculated over the full length of the reference nucleotide sequence and that gaps in homology of up to 5% of the total number of nucleotides in the reference sequence are allowed.

[0053] Preferred are nucleic acid molecules having sequences at least 90%, 95%, 96%, 97%, 98% or 99% identical to the nucleic acid sequence of an *E. coli* J96 PAI ORF that encode a functional polypeptide. By a "functional polypeptide" is intended a polypeptide exhibiting activity similar, but not necessarily identical, to an activity of the protein encoded by the *E. coli* J96 PAI ORF. For example, the *E. coli* ORF [Contig ID 84, ORF ID 3 (84/3)] encodes a hemolysin. Thus, a functional polypeptide encoded by a nucleic acid molecule having a nucleotide sequence, for example, 95% identical to the nucleotide sequence of 84/3, will also possess hemolytic activity. As the skilled artisan will appreciate, assays for determining whether a particular polypeptide is functional will depend on which ORF is used as the reference sequence. Depending on the reference ORF, the assay chosen for measuring polypeptide activity will be readily apparent in light of the role categories provided in Tables 1, 3, 5 and 6.

[0054] Of course, due to the degeneracy of the genetic code, one of ordinary skill in the art will immediately recognize that a large number of the nucleic acid molecules having a sequence at least 90%, 95%, 96%, 97%, 98%, or 99% identical to the nucleic acid sequence of a reference ORF will encode a functional polypeptide. In fact, since degenerate variants all encode the same amino acid sequence, this will be clear to the skilled artisan even without performing a comparison assay for protein activity. It will be further recognized in the art that, for such nucleic acid molecules that are not degenerate

variants, a reasonable number will also encode a functional polypeptide. This is because the skilled artisan is fully aware of amino acid substitutions that are either less likely or not likely to significantly affect protein function (e.g., replacing one aliphatic amino acid with a second aliphatic amino acid).

[0055] For example, guidance concerning how to make phenotypically silent amino acid substitutions is provided in Bowie, J. U. *et al.,* "Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitutions," *Science 247:*1306-1310 (1990), wherein the authors indicate that there are two main approaches for studying the tolerance of an amino acid sequence to change. The first method relies on the process of evolution, in which mutations are either accepted or rejected by natural selection. The second approach uses genetic engineering to introduce amino acid changes at specific positions of a cloned gene and selections or screens to identify sequences that maintain functionality. As the authors state, these studies have revealed that proteins are surprisingly tolerant of amino acid substitutions. The authors further indicate which amino acid changes are likely to be permissive at a certain position of the protein. For example, most buried amino acid residues require nonpolar side chains, whereas few features of surface side chains are generally conserved. Other such phenotypically silent substitutions are described in Bowie, J.U. *et al., supra,* and the references cited therein.

[0056] The present invention is further directed to fragments of the isolated nucleic acid molecules described herein. By a fragment of an isolated nucleic acid molecule having the nucleotide sequence of an *E. coli* J96 PAI ORF is intended fragments at least about 15 nt, and more preferably at least about 20 nt, still more preferably at least about 30 nt, and even more preferably, at least about 40 nt in length that are useful as diagnostic probes and primers as discussed herein. Of course, larger fragments 50-500 nt in length are also useful according to the present invention as are fragments corresponding to most, if not all, of the nucleotide sequence of an *E. coli* J96 PAI ORF. By a fragment at least 20 nt in length, for example, is intended fragments that include 20 or more contiguous bases from the nucleotide sequence of an *E. coli* J96 PAI ORF. Since *E. coli* ORFs are listed in Tables 1 through 6 and the sequences of the ORFs have been provided within the contig sequences of SEQ ID NOs: 1 through 142, generating such DNA fragments would be routine to the skilled artisan. For example, restriction endonuclease cleavage or shearing by sonication could easily be used to generate fragments of various sizes from the PAI DNA that is incorporated into the deposited pools of cosmid clones. Alternatively, such fragments could be generated synthetically.

[0057] Preferred nucleic acid fragments of the present invention include nucleic acid molecules encoding epitope-bearing portions of an *E. coli* J96 PAI protein. Methods for determining such epitope-bearing portions are described in detail below.

[0058] In another aspect, the invention provides an isolated nucleic acid molecule comprising a polynucleotide that hybridizes under stringent hybridization conditions to a portion of the polynucleotide in a nucleic acid molecule of the invention described above, for instance, an ORF described in Tables 1 through 6, preferably an ORF described in Tables 1, 2, 3 or 4. By "stringent hybridization conditions" is intended overnight incubation at 42°C in a solution comprising: 50% formamide, 5 x SSC (750 mM NaCl, 75mM trisodium citrate), 50 mM sodium phosphate (pH 7.6), 5 x Denhardt's solution, 10% dextran sulfate, and 20 g/ml denatured, sheared salmon sperm DNA, followed by washing the filters in 0.1 x SSC at about 65°C.

[0059] By a polynucleotide that hybridizes to a "portion" of a polynucleotide is intended a polynucleotide (either DNA or RNA) hybridizing to at least about 15 nucleotides (nt), and more preferably at least about 20 nt, still more preferably at least about 30 nt, and even more preferably about 30-70 nt of the reference polynucleotide. These are useful as diagnostic probes and primers as discussed above and in more detail below.

[0060] Of course, polynucleotides hybridizing to a larger portion of the reference polynucleotide (e.g., a *E. coli* ORF), for instance, a portion 50-500 nt in length, or even to the entire length of the reference polynucleotide, are also useful as probes according to the present invention, as are polynucleotides corresponding to most, if not all, of an *E. coli* J96 PAI ORF.

[0061] By "expression modulating fragment" (EMF), is intended a series of nucleotides that modulate the expression of an operably linked ORF or EMF. A sequence is said to "modulate the expression of an operably linked sequence" when the expression of the sequence is altered by the presence of the EMF. EMFs include, but are not limited to, promoters, and promoter modulating sequences (inducible elements). One class of EMFs are fragments that induce the expression of an operably linked ORF in response to a specific regulatory factor or physiological event. EMF sequences can be identified within the *E. coli* genome by their proximity to the ORFs described in Tables 1 through 6. An intergenic segment, or a fragment of the intergenic segment, from about 10 to 200 nucleotides in length, taken 5' from any one of the ORFs of Tables 1 through 6 will modulate the expression of an operably linked 3' ORF in a fashion similar to that found with the naturally linked ORF sequence. As used herein, an "intergenic segment" refers to the fragments of the *E. coli* J96 PAI subgenome that are between two ORF(s) herein

described. Alternatively, EMFs can be identified using known EMFs as a target sequence or target motif in the computer-based systems of the present invention.

[0062] The presence and activity of an EMF can be confirmed using an EMF trap vector. An EMF trap vector contains a cloning site 5' to a marker sequence. A marker sequence encodes an identifiable phenotype, such as antibiotic resistance or a complementing nutrition auxotrophic factor, which can be identified or assayed when the EMF trap vector is placed within an appropriate host under appropriate conditions. As described above, an EMF will modulate the expression of an operably linked marker sequence. A more detailed discussion of various marker sequences is provided below.

[0063] A sequence that is suspected as being an EMF is cloned in all three reading frames in one or more restriction sites upstream from the marker sequence in the EMF trap vector. The vector is then transformed into an appropriate host using known procedures and the phenotype of the transformed host in examined under appropriate conditions. As described above, an EMF will modulate the expression of an operably linked marker sequence.

[0064] By a "diagnostic fragment" (DF), is intended a series of nucleotides that selectively hybridize to *E. coli* sequences. DFs can be readily identified by identifying unique sequences within the *E. coli* J96 PAI subgenome, or by generating and testing probes or amplification primers consisting of the DF sequence in an appropriate diagnostic format for amplification or hybridization selectivity.

[0065] Each of the ORFs of the *E. coli* J96 PAI subgenome disclosed in Tables 1 through 4, and the EMF found 5' to the ORF, can be used in numerous ways as polynucleotide reagents. The sequences can be used as diagnostic probes or diagnostic amplification primers to detect the presence of uropathogenic *E. coli* in a sample. This is especially the case with the fragments or ORFs of Table 2 and 4 which will be highly selective for uropathogenic *E. coli* J96, and perhaps other uropathogenic or extraintestinal strains that include one or more PAIs.

[0066] In addition, the fragments of the present invention, as broadly described, can be used to control gene expression through triple helix formation or antisense DNA or RNA, both of which methods are based on the binding of a polynucleotide sequence to DNA or RNA. Polynucleotides suitable for use in these methods are usually 20 to 40 bases in length and are designed to be complementary to a region of the gene involved in transcription (triple helix - see Lee *et al.*, *Nucl. Acids Res. 6*:3073 (1979); Cooney *et al.*, *Science 241*:456 (1988); and Dervan *et al.*, *Science 251*:1360 (1991)) or to the mRNA itself (antisense - Okano, *J. Neurochem. 56*:560 (1991); *Oligodeoxynucleotides as Antisense Inhibitors of Gene Expression*, CRC Press, Boca Raton, FL (1988)).

[0067] Triple helix- formation optimally results in a shut-off of RNA transcription from DNA, while antisense RNA hybridization blocks translation of an mRNA molecule into polypeptide. Both techniques have been demonstrated to be effective in model systems. Information contained in the sequences of the present invention is necessary for the design of an antisense or triple helix oligonucleotide.

Vectors and Host Cells

[0068] The present invention further provides recombinant constructs comprising one or more fragments of the *E. coli* J96 PAIs. The recombinant constructs of the present invention comprise a vector, such as a plasmid or viral vector, into which, for example, an *E. coli* J96 PAI ORF is inserted. The vector may further comprise regulatory sequences, including for example, a promoter, operably linked to the ORF. For vectors comprising the EMFs of the present invention, the vector may further comprise a marker sequence or heterologous ORF operably linked to the EMF. Large numbers of suitable vectors and promoters are known to those of skill in the art and are commercially available for generating the recombinant constructs of the present invention. The following vectors are provided by way of example. Bacterial: pBs, phagescript, PsiX174, pBluescript SK, pBs KS, pNH8a, pNH16a, pNH18a, pNH46a (Stratagene); pTrc99A, pKK223-3, pKK233-3, pDR540, pRIT5 (Pharmacia). Eukaryotic: pWLneo, pSV2cat, pOG44, pXT1, pSG (Stratagene) pSVK3, pBPV, pMSG, pSVL (Pharmacia).

[0069] Promoter regions can be selected from any desired gene using CAT (chloramphenicol transferase) vectors or other vectors with selectable markers. Two appropriate vectors are pKK232-8 and pCM7. Particular named bacterial promoters include lacI, lacZ, T3, T7, gpt, lambda P_R, and trc. Eukaryotic promoters include CMV immediate early, HSV thymidine kinase, early and late SV40, LTRs from retrovirus, and mouse metallothionein-I. Selection of the appropriate vector and promoter is well within the level of ordinary skill in the art.

[0070] The present invention further provides host cells containing any one of the isolated fragments (preferably an ORF) of the *E. coli* J96 PAIs described herein. The host cell can be a higher eukaryotic host cell, such as a mammalian cell, a lower eukaryotic host cell, such as a yeast cell, or the host cell can be a procaryotic cell, such as a bacterial cell. Introduction of the recombinant construct into the host cell can be effected by calcium phosphate transfection, DEAE, dextran mediated transfection, or electroporation (Davis, L. *et al.*, *Basic Methods in Molecular Biology* (1986)). Host cells containing, for example, an *E. coli* J96 PAI ORF can be used conventionally to produce the encoded protein.

Polypeptides and Fragments

[0071] The invention further provides isolated polypeptides having the amino acid sequence encoded by an *E coli* PAI ORF described in Tables 1 through 6, preferably Tables 1 through 4, or a peptide or polypeptide comprising a portion of the above polypeptides. The terms "peptide" and "oligopeptide" are considered synonymous (as is commonly recognized) and each term can be used interchangeably as the context requires to indicate a chain of at least to amino acids coupled by peptidyl linkages. The word "polypeptide" is used herein for chains containing more than ten amino acid residues. All oligopeptide and polypeptide formulas or sequences herein are written from left to right and in the direction from amino terminus to carboxy terminus.

[0072] It will be recognized in the art that some amino acid sequences of *E. coli* polypeptides can be varied without significant effect of the structure or function of the protein. If such differences in sequence are contemplated, it should be remembered that there will be critical areas on the protein which determine activity. In general, it is possible to replace residues that form the tertiary structure, provided that residues performing a similar function are used. In other instances, the type of residue may be completely unimportant if the alteration occurs at a non-critical region of the protein.

[0073] Thus, the invention further includes variations of polypeptides encoded for by ORFs listed in Tables 1 through 6 which show substantial pathogenic activity or which include regions of particular *E. coli* PAI proteins such as the protein portions discussed below. Such mutants include deletions, insertions, inversions, repeats, and type substitutions (for example, substituting one hydrophilic residue for another, but not strongly hydrophilic for strongly hydrophobic as a rule). Small changes or such "neutral" amino acid substitutions will generally have little effect on activity.

[0074] Typically seen as conservative substitutions are the replacements, one for another, among the aliphatic amino acids Ala, Val, Leu and Ile; interchange of the hydroxyl residues Ser and Thr, exchange of the acidic residues Asp and Glu, substitution between the amide residues Asn and Gln, exchange of the basic residues Lys and Arg and replacements among the aromatic residues Phe, Tyr.

[0075] As indicated in detail above, further guidance concerning which amino acid changes are likely to be phenotypically silent (i.e., are not likely to have a significant deleterious effect on a function) can be found in Bowie, J.U., *et al.*, "Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitutions," *Science 247:*1306-1310 (1990).

[0076] Thus, the fragment, derivative or analog of a polypeptide encoded by an ORF described in one of Tables 1 through 6, may be (i) one in which one or more of the amino acid residues are substituted with a conserved or non-conserved amino acid residue (preferably a conserved amino acid residue) and such substituted amino acid residue may or may not be one encoded by the genetic code, or (ii) one in which one or more of the amino acid residues includes a substituent group, or (iii) one in which the mature polypeptide is fused with another compound, such as a compound to increase the half-life of the polypeptide (for example, polyethylene glycol), or (iv) one in which the additional amino acids are fused to the mature polypeptide, such as an IgG Fc fusion region peptide or leader or secretory sequence or a sequence which is employed for purification of the mature polypeptide or a proprotein sequence. Such fragments, derivatives and analogs are deemed to be within the scope of those skilled in the art from the teachings herein.

[0077] Of particular interest are substitutions of charged amino acids with another charged amino acid and with neutral or negatively charged amino acids. The latter results in proteins with reduced positive charge to improve the characteristics of said proteins. The prevention of aggregation is highly desirable. Aggregation of proteins not only results in a loss of activity but can also be problematic when preparing pharmaceutical formulations, because they can be immunogenic. (Pinckard *et al., Clin Exp. Immunol. 2*:331-340 (1967); Robbins *et al., Diabetes 36*:838-845 (1987); Cleland *et al. Crit. Rev. Therapeutic Drug Carrier Systems 10*:307-377 (1993)).

[0078] The replacement of amino acids can also change the selectivity of binding to cell surface receptors. Ostade *et al., Nature 361*:266-268 (1993) describes certain mutations resulting in selective binding of TNF-Â to only one of the two known types of TNF receptors. Thus, proteins encoded for by the ORFs listed in Tables 1, 2, 3, 4, 5, or 6, and that bind to a cell surface receptor, may include one or more amino acid substitutions, deletions or additions, either from natural mutations or human manipulation.

[0079] As indicated, changes are preferably of a minor nature, such as conservative amino acid substitutions that do not significantly affect the folding or activity of the protein (see Table 7).

TABLE 7. Conservative Amino Acid Substitutions

Aromatic	Phenylalanine Tryptophan Tyrosine
Hydrophobic	Leucine Isoleucine Valine
Polar	Glutamine Asparagine
Basic	Arginine Lysine Histidine
Acidic	Aspartic Acid Glutamic Acid
Small	Alanine Serine Threonine Methionine
Glycine	

[0080] Amino acids in the proteins encoded by ORFs of the present invention that are essential for function can be identified by methods known in the art, such as site-directed mutagenesis or alanine-scanning mutagenesis (Cunningham and Wells, *Science 244*:1081-1085 (1989)). The latter procedure introduces single alanine mutations at every residue in the molecule. The resulting mutant molecules are then tested for biological activity such as receptor binding or *in vitro*, or *in vitro* proliferative activity. Sites that are critical for ligand-receptor binding can also be determined by structural analysis such as crystallization, nuclear magnetic resonance or photoaffinity labeling (Smith *et al., J. Mol. Biol. 224*:899-904 (1992) and de Vos *et al. Science 255*:306-312 (1992)).

[0081] The polypeptides of the present invention are preferably provided in an isolated form, and preferably are substantially purified. A recombinantly produced version of the polypeptides can be substantially purified by the one-step method described in Smith and Johnson, *Gene 67*:31-40 (1988).

[0082] The polypeptides of the present invention include the polypeptide encoded by the ORFs listed in Tables 1-6, preferably Tables 1-4, as well as polypeptides which have at least 90% similarity, more preferably at least 95% similarity, and still more preferably at least 96%, 97%, 98% or 99% similarity to those described above, and also include portions of such polypeptides with at least 30 amino acids and more preferably at least 50 amino acids.

[0083] By "% similarity" for two polypeptides is intended a similarity score produced by comparing the amino acid sequences of the two polypeptides using the Bestfit program (Wisconsin Sequence Analysis Package, Version 8 for Unix, Genetics Computer Group, University Research Park, 575 Science Drive, Madison, WI 53711) and the default settings for determining similarity. Bestfit uses the local homology algorithm of Smith and Waterman *(Advances in Applied Mathematics 2*:482-489, 1981) to find the best segment of similarity between two sequences.

[0084] By a polypeptide having an amino acid sequence at least, for example, 95% "identical" to a reference amino acid sequence of a polypeptide is intended that the amino acid sequence of the polypeptide is identical to the reference sequence except that the polypeptide sequence may include up to five amino acid alterations per each 100 amino acids of the reference amino acid of said polypeptide. In other words, to obtain a polypeptide having an amino acid sequence at least 95% identical to a reference amino acid sequence, up to 5% of the amino acid residues in the reference sequence may be deleted or substituted with another amino acid, or a number of amino acids up to 5% of the total amino acid residues in the reference sequence may be inserted into the reference sequence. These alterations of the reference sequence may occur at the amino or carboxy terminal positions of the reference amino acid sequence or anywhere between those terminal positions, interspersed either individually among residues in the reference sequence or in one or more contiguous groups within the reference sequence.

[0085] As a practical matter, whether any particular polypeptide is at least 90%, 95%, 96%, 97%, 98% or 99% identical to, for instance, the amino acid sequence encoded by the ORFs listed in Tables 1, 2, 3, 4, 5, or 6 can be determined conventionally using known computer programs such the Bestfit program (Wisconsin Sequence Analysis Package, Version 8 for Unix, Genetics Computer Group, University Research Park, 575 Science Drive, Madison, WI 53711. When using Bestfit or any other sequence alignment program to determine whether a particular sequence is, for instance, 95% identical to a reference sequence according to the present invention, the parameters are set, of course, such that the percentage of identity is calculated over the full length of the reference amino acid sequence and that gaps in homology of up to 5% of the total number of amino acid residues in the reference sequence are allowed.

[0086] The polypeptide of the present invention could be used as a molecular weight marker on SDS-PAGE gels or on molecular sieve gel filtration columns using methods well known to those of skill in the art.

[0087] As described in detail below, the polypeptides of the present invention can also be used to raise polyclonal and monoclonal antibodies, which are useful in assays for

detecting pathogenic protein expression as described below or as agonists and antagonists capable of enhancing or inhibiting protein function of important proteins encoded by the ORFs of the present invention. Further, such polypeptides can be used in the yeast two-hybrid system to "capture" protein binding proteins which are also candidate agonist and antagonist according to the present invention. The yeast two hybrid system is described in Fields and Song, *Nature 340:*245-246 (1989).

[0088] In another aspect, the invention provides a peptide or polypeptide comprising an epitope-bearing portion of a polypeptide of the invention. The epitope of this polypeptide portion is an immunogenic or antigenic epitope of a polypeptide of the invention. An "immunogenic epitope" is defined as a part of a protein that elicits an antibody response when the whole protein is the immunogen. These immunogenic epitopes are believed to be confined to a few loci on the molecule. On the other hand, a region of a protein molecule to which an antibody can bind is defined as an "antigenic epitope." The number of immunogenic epitopes of a protein generally is less than the number of antigenic epitopes. See, for instance, Geysen *et al., Proc. Natl. Acad. Sci. USA 81:*3998- 4002 (1983).

[0089] As to the selection of peptides or polypeptides bearing an antigenic epitope (i.e., that contain a region of a protein molecule to which an antibody can bind), it is well known in that art that relatively short synthetic peptides that mimic part of a protein sequence are routinely capable of eliciting an antiserum that reacts with the partially mimicked protein. See, for instance, Sutcliffe, J. G., Shinnick, T. M., Green, N. and Learner, R.A. (1983) Antibodies that react with predetermined sites on proteins. *Science 219:*660-666. Peptides capable of eliciting protein-reactive sera are frequently represented in the primary sequence of a protein, can be characterized by a set of simple chemical rules, and are confined neither to immunodominant regions of intact proteins (i.e., immunogenic epitopes) nor to the amino or carboxyl terminals. Peptides that are extremely hydrophobic and those of six or fewer residues generally are ineffective at inducing antibodies that bind to the mimicked protein; longer, peptides, especially those containing proline residues, usually are effective. Sutcliffe *et al., supra,* at 661. For instance, 18 of 20 peptides designed according to these guidelines, containing 8-39 residues covering 75% of the sequence of the influenza virus hemagglutinin HA1 polypeptide chain, induced antibodies that reacted with the HA1 protein or intact virus; and 12/12 peptides from the MuLV polymerase and 18/18 from the rabies glycoprotein induced antibodies that precipitated the respective proteins.

[0090] Antigenic epitope-bearing peptides and polypeptides of the invention are therefore useful to raise antibodies, including monoclonal antibodies that bind specifically

to a polypeptide of the invention. Thus, a high proportion of hybridomas obtained by fusion of spleen cells from donors immunized with an antigen epitope-bearing peptide generally secrete antibody reactive with the native protein. Sutcliffe *et al.*, *supra*, at 663. The antibodies raised by antigenic epitope-bearing peptides or polypeptides are useful to detect the mimicked protein, and antibodies to different peptides may be used for tracking the fate of various regions of a protein precursor which undergoes post-translational processing. The peptides and anti-peptide antibodies may be used in a variety of qualitative or quantitative assays for the mimicked protein, for instance in competition assays since it has been shown that even short peptides (e.g., about 9 amino acids) can bind and displace the larger peptides in immunoprecipitation assays. See, for instance, Wilson *et al.*, *Cell 37*:767-778 (1984) at 777. The anti- peptide antibodies of the invention also are useful for purification of the mimicked protein, for instance, by adsorption chromatography using methods well known in the art.

[0091] Antigenic epitope-bearing peptides and polypeptides of the invention designed according to the above guidelines preferably contain a sequence of at least seven, more preferably at least nine and most preferably between about 15 to about 30 amino acids contained within the amino acid sequence of a polypeptide of the invention. However, peptides or polypeptides comprising a larger portion of an amino acid sequence of a polypeptide of the invention, containing about 30 to about 50 amino acids, or any length up to and including the entire amino acid sequence of a polypeptide of the invention, also are considered epitope-bearing peptides or polypeptides of the invention and also are useful for inducing antibodies that react with the mimicked protein. Preferably, the amino acid sequence of the epitope-bearing peptide is selected to provide substantial solubility in aqueous solvents (i.e., the sequence includes relatively hydrophilic residues and highly hydrophobic sequences are preferably avoided); and sequences containing proline residues are particularly preferred.

[0092] The epitope-bearing peptides and polypeptides of the invention may be produced by any conventional means for making peptides or polypeptides including recombinant means using nucleic acid molecules of the invention. For instance, a short epitope-bearing amino acid sequence may be fused to a larger polypeptide, which acts as a carrier during recombinant production and purification, as well as during immunization to produce anti-peptide antibodies. Epitope-bearing peptides also may be synthesized using known methods of chemical synthesis. For instance, Houghten has described a simple method for synthesis of large numbers of peptides, such as 10-20 mg of 248 different 13 residue peptides representing single amino acid variants of a segment of the HA1 polypeptide which were prepared and characterized (by ELISA-type binding studies) in

less than four weeks. Houghten, R. A. (1985) General method for the rapid solid-phase synthesis of large numbers of peptides: specificity of antigen-antibody interaction at the level of individual amino acids. *Proc. Natl. Acad. Sci. USA 82*:5131-5135. This "Simultaneous Multiple Peptide Synthesis (SMPS)" process is further described in U.S. Patent No. 4,631,211 to Houghten *et al.* (1986). In this procedure the individual resins for the solid-phase synthesis of various peptides are contained in separate solvent-permeable packets, enabling the optimal use of the many identical repetitive steps involved in solid-phase methods. A completely manual procedure allows 500-1000 or more syntheses to be conducted simultaneously. Houghten *et al.*, *supra*, at 5134.

[0093] Epitope-bearing peptides and polypeptides of the invention are used to induce antibodies according to methods well known in the art. See, for instance, Sutcliffe *et al.*, *supra*; Wilson *et al.*, *supra*; Chow, M. *et al.*, *Proc. Natl. Acad. Sci. USA 82*:910-914; and Bittle, F. J. *et al.*, *J. Gen. Virol. 66*:2347-2354 (1985). Generally, animals may be immunized with free peptide; however, anti-peptide antibody titer may be boosted by coupling of the peptide to a macromolecular carrier, such as keyhole limpet hemacyanin (KLH) or tetanus toxoid. For instance, peptides containing cysteine may be coupled to carrier using a linker such as m-maleimidobenzoyl-N-hydroxysuccinimide ester (MBS), while other peptides may be coupled to carrier using a more general linking agent such as glutaraldehyde. Animals such as rabbits, rats and mice are immunized with either free or carrier-coupled peptides, for instance, by intraperitoneal and/or intradermal injection of emulsions containing about 100 μg peptide or carrier protein and Freund's adjuvant. Several booster injections may be needed, for instance, at intervals of about two weeks, to provide a useful titer of anti-peptide antibody, which can be detected, for example, by ELISA assay using free peptide adsorbed to a solid surface. The titer of anti-peptide antibodies in serum from an immunized animal may be increased by selection of anti-peptide antibodies, for instance, by adsorption to the peptide on a solid support and elution of the selected antibodies according to methods well known in the art.

[0094] Immunogenic epitope-bearing peptides of the invention, i.e., those parts of a protein that elicit an antibody response when the whole protein is the immunogen, are identified according to methods known in the art. For instance, Geysen *et al.*, *supra*, discloses a procedure for rapid concurrent synthesis on solid supports of hundreds of peptides of sufficient purity to react in an enzyme-linked immunosorbent assay. Interaction of synthesized peptides with antibodies is then easily detected without removing them from the support. In this manner a peptide bearing an immunogenic epitope of a desired protein may be identified routinely by one of ordinary skill in the art. For instance, the immunologically important epitope in the coat protein of foot-and-mouth

disease virus was located by Geysen *et al. supra* with a resolution of seven amino acids by synthesis of an overlapping set of all 208 possible hexapeptides covering the entire 213 amino acid sequence of the protein. Then, a complete replacement set of peptides in which all 20 amino acids were substituted in turn at every position within the epitope were synthesized, and the particular amino acids conferring specificity for the reaction with antibody were determined. Thus, peptide analogs of the epitope-bearing peptides of the invention can be made routinely by this method. U.S. Patent No. 4,708,781 to Geysen (1987) further describes this method of identifying a peptide bearing an immunogenic epitope of a desired protein.

[0095] Further still, U.S. Patent No. 5,194,392 to Geysen (1990) describes a general method of detecting or determining the sequence of monomers (amino acids or other compounds) which is a topological equivalent of the epitope (i.e., a "mimotope") which is complementary to a particular paratope (antigen binding site) of an antibody of interest. More generally, U.S. Patent No. 4,433,092 to Geysen (1989) describes a method of detecting or determining a sequence of monomers which is a topographical equivalent of a ligand which is complementary to the ligand binding site of a particular receptor of interest. Similarly, U.S. Patent No. 5,480,971 to Houghten, R. A. *et al.* (1996) on Peralkylated Oligopeptide Mixtures discloses linear C -C -alkyl peralkylated oligopeptides and sets and libraries of such peptides, as well as methods for using such oligopeptide sets and libraries for determining the sequence of a peralkylated oligopeptide that preferentially binds to an acceptor molecule of interest. Thus, non-peptide analogs of the epitope-bearing peptides of the invention also can be made routinely by these methods.

[0096] The entire disclosure of each document cited in this section on "Polypeptides and Peptides" is hereby incorporated herein by reference.

[0097] As one of skill in the art will appreciate, *E. coli* PAI polypeptides of the present invention and the epitope-bearing fragments thereof described above can be combined with parts of the constant domain of immunoglobulins (IgG), resulting in chimeric polypeptides. These fusion proteins facilitate purification and show an increased half-life *in vivo.* This has been shown, e.g., for chimeric proteins consisting of the first two domains of the human CD4-polypeptide and various domains of the constant regions of the heavy or light chains of mammalian immunoglobulins (EP A 394,827; Traunecker *et al., Nature 331*:84- 86 (1988)). Fusion proteins that have a disulfide-linked dimeric structure due to the IgG part can also be more efficient in binding and neutralizing other molecules than the monomeric *E. coli* J96 PAI proteins or protein fragments alone (Fountoulakis *et al., J. Biochem 270*:3958-3964 (1995)).

Vaccines

[0098] In another embodiment, the present invention relates to a vaccine, preferably in unit dosage form, comprising one or more *E. coli* J96 PAI antigens together with a pharmaceutically acceptable diluent, carrier, or excipient, wherein the one or more antigens are present in an amount effective to elicit a protective immune response in an animal to pathogenic *E. coli*. Antigens of *E. coli* J96 PAI IV and V may be obtained from polypeptides encoded for by the ORFs listed in Tables 1-6, particularly Tables 1-4, using methods well known in the art.

[0099] In a preferred embodiment, the antigens are *E. coli* J96 PAI IV or PAI V proteins that are present on the surface of pathogenic *E. coli*. In another preferred embodiment, the pathogenic *E. coli* J96 PAI IV or PAI V protein-antigen is conjugated to an *E. coli* capsular polysaccharide (CP), particularly to capsular polypeptides that are more prevalent in pathogenic strains, to produce a double vaccine. CPs, in general, may be prepared or synthesized as described in Schneerson *et al. J. Exp. Med. 152:*361-376 (1980); Marburg *et al. J. Am. Chem. Soc.* 108:5282 (1986); Jennings *et al., J. Immunol. 127:*1011-1018 (1981); and Beuvery *et al., Infect. Immunol. 40:*39-45 (1983). In a further preferred embodiment, the present invention relates to a method of preparing a polysaccharide conjugate comprising: obtaining the above-described *E. coli* J96 PAI antigen; obtaining a CP or fragment from pathogenic *E. coli*; and conjugating the antigen to the CP or CP fragment.

[0100] In a preferred embodiment, the animal to be protected is selected from the group consisting of humans, horses, deer, cattle, pigs, sheep, dogs, and chickens. In a more preferred embodiment, the animal is a human or a dog.

[0101] In a further embodiment, the present invention relates to a prophylactic method whereby the incidence of pathogenic *E. coli*-induced symptoms are decreased in an animal, comprising administering to the animal the above-described vaccine, wherein the vaccine is administered in an amount effective to elicit protective antibodies in an animal to pathogenic *E. coli*. This vaccination method is contemplated to be useful in protecting against severe diarrhea (pathogenic intestinal *E. coli* strains), urinary tract infections (uropathogenic *E. coli*) and infections of the brain (extraintestinal *E. coli*). The vaccine of the invention is used in an effective amount depending on the route of administration. Although intra-nasal, subcutaneous or intramuscular routes of administration are preferred, the vaccine of the present invention can also be administered by an oral, intraperitoneal or intravenous route. One skilled in the art will appreciate that the amounts to be administered for any particular treatment protocol can be readily determined without

undue experimentation. Suitable amounts are within the range of 2 micrograms of the protein per kg body weight to 100 micrograms per kg body weight.

[0102] The vaccine can be delivered through a vector such as BCG. The vaccine can also be delivered as naked DNA coding for target antigens.

[0103] The vaccine of the present invention may be employed in such dosage forms as capsules, liquid solutions, suspensions or elixirs for oral administration, or sterile liquid forms such as solutions or suspensions. Any inert carrier is preferably used, such as saline, phosphate-buffered saline, or any such carrier in which the vaccine has suitable solubility properties. The vaccines may be in the form of single dose preparations or in multi-dose flasks which can be used for mass vaccination programs. Reference is made to Remington's *Pharmaceutical Sciences*, Mack Publishing Co., Easton, PA, Osol (ed.) (1980); and *New Trends and Developments in Vaccines*, Voller *et al.* (eds.), University Park Press, Baltimore, MD (1978), for methods of preparing and using vaccines.

[0104] The vaccines of the present invention may further comprise adjuvants which enhance production of antibodies and immune cells. Such adjuvants include, but are not limited to, various oil formulations such as Freund's complete adjuvant (CFA), the dipeptide known as MDP, saponins (ex. *Quillajasaponin* fraction QA-21, U.S. Patent No. 5,047,540), aluminum hydroxide, or lymphatic cytokines. Freund's adjuvant is an emulsion of mineral oil and water which is mixed with the immunogenic substance. Although Freund's adjuvant is powerful, it is usually not administered to humans. Instead, the adjuvant alum (aluminum hydroxide) may be used for administration to a human. Vaccine may be absorbed onto the aluminum hydroxide from which it is slowly released after injection. The vaccine may also be encapsulated within liposomes according to Fullerton, U.S. Patent No. 4,235,877.

Protein Function

[0105] Each ORF described in Tables 1 and 3 possesses a biological role similar to the role associated with the identified homologous protein. This allows the skilled artisan to determine a function for each identified coding sequence. For example, a partial list of the *E. coli* protein functions provided in Tables 1 and 3 includes many of the functions associated with virulence of pathogenic bacterial strains. These include, but are not limited to adhesins, excretion pathway proteins, O-antigen/carbohydrate modification, cytotoxins and regulators. A more detailed description of several of these functions is provided in Example 1 below.

Diagnostic Assays

[0106] In another preferred embodiment, the present invention relates to a method of detecting pathogenic *E. coli* nucleic acid in a sample comprising:

(a) contacting the sample with the above-described nucleic acid probe, under conditions such that hybridization occurs, and

(b) detecting the presence of the probe bound to pathogenic *E. coli* nucleic acid.

[0107] In another preferred embodiment, the present invention relates to a diagnostic kit for detecting the presence of pathogenic *E. coli* nucleic acid in a sample comprising at least one container means having disposed therein the above-described nucleic acid probe.

[0108] In another preferred embodiment, the present invention relates to a diagnostic kit for detecting the presence of pathogenic *E. coli* antigens in a sample comprising at least one container means having disposed therein the above-described antibodies.

[0109] In another preferred embodiment, the present invention relates to a diagnostic kit for detecting the presence of antibodies to pathogenic *E. coli* antigens in a sample comprising at least one container means having disposed therein the above-described antigens.

[0110] The present invention provides methods to identify the expression of an ORF of the present invention, or homolog thereof, in a test sample, using one of the antibodies of the present invention. Such methods involve incubating a test sample with one or more of the antibodies of the present invention and assaying for binding of the antibodies to components within the test sample.

[0111] In a further embodiment, the present invention relates to a method for identifying pathogenic *E. coli* in an animal comprising analyzing tissue or body fluid from the animal for a nucleic acid, protein, polypeptide-antigen or antibody specific to one of the ORFs described in Tables 1-4 herein from *E. coli* J96 PAI IV or V. Analysis of nucleic acid specific to pathogenic *E. coli* can be by PCR techniques or hybridization techniques (cf. *Molecular Cloning: A Laboratory Manual, second edition*, edited by Sambrook, Fritsch, & Maniatis, Cold Spring Harbor Laboratory, 1989; Eremeeva *et al., J. Clin. Microbiol. 32:*803-810 (1994) which describes differentiation among spotted fever group *Rickettsiae* species by analysis of restriction fragment length polymorphism of PCR-amplified DNA).

[0112] Proteins or antibodies specific to pathogenic *E. coli* may be identified as described in *Molecular Cloning: A Laboratory Manual, second edition*, Sambrook *et al.*, eds., Cold Spring Harbor Laboratory (1989). More specifically, antibodies may be raised to *E. coli* J96 PAI proteins as generally described in *Antibodies: A Laboratory Manual*,

Harlow and Lane, eds., Cold Spring Harbor Laboratory (1988). *E. coli* J96 PAI-specific antibodies can also be obtained from infected animals (Mather, T. *et al.*, *JAMA 205*:186-188 (1994)).

[0113] In another embodiment, the present invention relates to an antibody having binding affinity specifically to an *E. coli* J96 PAI antigen as described above. The *E. coli* J96 PAI antigens of the present invention can be used to produce antibodies or hybridomas. One skilled in the art will recognize that if an antibody is desired, a peptide can be generated as described herein and used as an immunogen. The antibodies of the present invention include monoclonal and polyclonal antibodies, as well as fragments of these antibodies. The invention further includes single chain antibodies. Antibody fragments which contain the idiotype of the molecule can be generated by known techniques, for example, such fragments include but are not limited to: the F(ab) fragment; the Fab fragments, Fab fragments, and Fv fragments.

[0114] Of special interest to the present invention are antibodies to pathogenic *E. coli* antigens which are produced in humans, or are "humanized" (i.e. non-immunogenic in a human) by recombinant or other technology. Humanized antibodies may be produced, for example by replacing an immunogenic portion of an antibody with a corresponding, but non-immunogenic portion (i.e. chimeric antibodies) (Robinson, R.R. *et al.*, International Patent Publication PCT/US86/02269; Akira, K. *et al.*, European Patent Application 184,187; Taniguchi, M., European Patent Application 171,496; Morrison, S.L. *et al.*, European Patent Application 173,494; Neuberger, M.S. et al., PCT Application WO 86/01533; Cabilly, S. *et al.*, European Patent Application 125,023; Better, M. *et al.*, *Science 240*:1041-1043 (1988); Liu, A.Y. *et al., Proc. Natl. Acad. Sci. USA 84*:3439-3443 (1987); Liu, A.Y. *et al., J. Immunol. 139*:3521-3526 (1987); Sun, L.K. *et al., Proc. Natl. Acad. Sci. USA 84*:214-218 (1987); Nishimura, Y. *et al., Canc. Res. 47*:999-1005 (1987); Wood, C.R. *et al., Nature 314*:446-449 (1985)); Shaw *et al., J. Natl. Cancer Inst. 80*:1553-1559 (1988). General reviews of "humanized" chimeric antibodies are provided by Morrison, S.L. (*Science, 229*:1202-1207 (1985)) and by Oi, V.T. *et al., BioTechniques 4*:214 (1986)). Suitable "humanized" antibodies can be alternatively produced by CDR or CEA substitution (Jones, P.T. *et al., Nature 321*:552-525 (1986); Verhoeyan *et al., Science 239*:1534 (1988); Beidler, C.B. *et al., J. Immunol. 141*:4053-4060 (1988)).

[0115] In another embodiment, the present invention relates to a hybridoma which produces the above-described monoclonal antibody. A hybridoma is an immortalized cell line which is capable of secreting a specific monoclonal antibody.

[0116] In general, techniques for preparing monoclonal antibodies and hybridomas are well known in the art (Campbell, "*Monoclonal Antibody Technology: Laboratory*

Techniques in Biochemistry and Molecular Biology," Elsevier Science Publishers, Amsterdam, The Netherlands (1984); St. Groth *et al.*, *J. Immunol. Methods* *35*:1-21 (1980)).

[0117] In another embodiment, the present invention relates to a method of detecting a pathogenic *E. coli* antigen in a sample, comprising: a) contacting the sample with an above-described antibody, under conditions such that immunocomplexes form, and b) detecting the presence of said antibody bound to the antigen. In detail, the methods comprise incubating a test sample with one or more of the antibodies of the present invention and assaying whether the antibody binds to the test sample.

[0118] Conditions for incubating an antibody with a test sample vary. Incubation conditions depend on the format employed in the assay, the detection methods employed, and the type and nature of the antibody used in the assay. One skilled in the art will recognize that any one of the commonly available immunological assay formats (such as radioimmunoassays, enzyme-linked immunosorbent assays, diffusion based Ouchterlony, or rocket immunofluorescent assays) can readily be adapted to employ the antibodies of the present invention. Examples of such assays can be found in Chard, *An Introduction to Radioimmunoassay and Related Techniques*, Elsevier Science Publishers, Amsterdam, The Netherlands (1986); Bullock *et al.*, *Techniques in Immunocytochemistry,* Academic Press, Orlando, FL Vol. 1 (1982), Vol. 2 (1983), Vol. 3 (1985); Tijssen, *Practice and Theory of Enzyme Immunoassays: Laboratory Techniques in Biochemistry and Molecular Biology,* Elsevier Science Publishers, Amsterdam, The Netherlands (1985); and *Antibodies: A Laboratory Manual*, Harlow and Lane, eds., Cold Spring Harbor Laboratory (1988).

[0119] The immunological assay test samples of the present invention include cells, protein or membrane extracts of cells, or biological fluids such as blood, serum, plasma, or urine. The test sample used in the above-described method will vary based on the assay format, nature of the detection method and the tissues, cells or extracts used as the sample to be assayed. Methods for preparing protein extracts or membrane extracts of cells are well known in the art and can be readily be adapted in order to obtain a sample which is capable with the system utilized.

[0120] In another embodiment, the present invention relates to a method of detecting the presence of antibodies to pathogenic *E. coli* in a sample, comprising: a) contacting the sample with an above-described antigen, under conditions such that immunocomplexes form, and b) detecting the presence of said antigen bound to the antibody. In detail, the methods comprise incubating a test sample with one or more of the antigens of the present invention and assaying whether the antigen binds to the test sample.

[0121] In another embodiment of the present invention, a kit is provided which contains all the necessary reagents to carry out the previously described methods of detection. The kit may comprise: i) a first container means containing an above-described antibody, and ii) second container means containing a conjugate comprising a binding partner of the antibody and a label. In another preferred embodiment, the kit further comprises one or more other containers comprising one or more of the following: wash reagents and reagents capable of detecting the presence of bound antibodies. Examples of detection reagents include, but are not limited to, labeled secondary antibodies, or in the alternative, if the primary antibody is labeled, the chromophoric, enzymatic, or antibody binding reagents which are capable of reacting with the labeled antibody. The compartmentalized kit may be as described above for nucleic acid probe kits.

[0122] One skilled in the art will readily recognize that the antibodies described in the present invention can readily be incorporated into one of the established kit formats which are well known in the art.

Screening Assay for Binding Agents

[0123] Using the isolated proteins described herein, the present invention further provides methods of obtaining and identifying agents that bind to a protein encoded by an *E. coli* J96 PAI ORF or to a fragment thereof.

[0124] The method involves:

(a) contacting an agent with an isolated protein encoded by a *E. coli* J96 PAI ORF, or an isolated fragment thereof; and

(b) determining whether the agent binds to said protein or said fragment.

[0125] The agents screened in the above assay can be, but are not limited to, peptides, carbohydrates, vitamin derivatives, or other pharmaceutical agents. The agents can be selected and screened at random or rationally selected or designed using protein modeling techniques. For random screening, agents such as peptides, carbohydrates, pharmaceutical agents and the like are selected at random and are assayed for their ability to bind to the protein encoded by an ORF of the present invention.

[0126] Alternatively, agents may be rationally selected or designed. As used herein, an agent is said to be "rationally selected or designed" when the agent is chosen based on the configuration of the particular protein. For example, one skilled in the art can readily adapt currently available procedures to generate peptides, pharmaceutical agents and the like capable of binding to a specific peptide sequence in order to generate rationally designed antipeptide ligands, for example see Hurby *et al.*, Application of Synthetic Peptides:

Antisense Peptides, In *Synthetic Peptides, A User's Guide*, W.H. Freeman, NY (1992), pp. 289-307, and Kaspczak *et al., Biochemistry 28*:9230-8 (1989).

[0127] In addition to the foregoing, one class of agents of the present invention, can be used to control gene expression through binding to one of the ORFs or EMFs of the present invention. As described above, such agents can be randomly screened or rationally designed and selected. Targeting the ORF or EMF allows a skilled artisan to design sequence specific or element specific agents, modulating the expression of either a single ORF or multiple ORFs that rely on the same EMF for expression control.

[0128] One class of DNA binding agents are those that contain nucleotide base residues that hybridize or form a triple helix by binding to DNA or RNA. Such agents can be based on the classic phosphodiester, ribonucleic acid backbone, or can be a variety of sulfhydryl or polymeric derivatives having base attachment capacity.

[0129] Agents suitable for use in these methods usually contain 20 to 40 bases and are designed to be complementary to a region of the gene involved in transcription (triple helix - see Lee *et al., Nucl. Acids Res. 6*:3073 (1979); Cooney *et al., Science 241*:456 (1988); and Dervan *et al., Science 251*: 1360 (1991)) or to the mRNA itself (antisense - Okano, *J. Neurochem. 56*:560 (1991); *Oligodeoxynucleotides as Antisense Inhibitors of Gene Expression*, CRC Press, Boca Raton, FL (1988)). Triple helix-formation optimally results in a shut-off of RNA transcription from DNA, while antisense RNA hybridization blocks translation of an mRNA molecule into polypeptide. Both techniques have been demonstrated to be effective in model systems. Information contained in the sequences of the present invention is necessary for the design of an antisense or triple helix oligonucleotide and other DNA binding agents.

Computer Related Embodiments

[0130] The nucleotide sequence provided in SEQ ID NOs: 1 through 142, representative fragments thereof, or nucleotide sequences at least 99.9% identical to the sequences provided in SEQ ID NOs: 1 through 142, can be "provided" in a variety of media to facilitate use thereof. As used herein, "provided" refers to a manufacture, other than an isolated nucleic acid molecule, that contains a nucleotide sequence of the present invention, i.e., the nucleotide sequence provided in SEQ ID NOs: 1 through 142, a representative fragment thereof, or a nucleotide sequence at least 99.9% identical to SEQ ID NOs: 1 through 142. Such a manufacture provides the *E. coli* J96 PAI subgenomes or a subset thereof (e.g., one or more *E. coli* J96 PAI open reading frame (ORF)) in a form that allows a skilled artisan to examine the manufacture using means not directly

applicable to examining the *E. coli* J96 PAI subgenome or a subset thereof as it exists in nature or in purified form.

[0131] In one application of this embodiment, one or more nucleotide sequences of the present invention can be recorded on computer readable media. As used herein, "computer readable media" refers to any medium that can be read and accessed directly by a computer. Such media include, but are not limited to: magnetic storage media, such as floppy discs, hard disc storage medium, and magnetic tape; optical storage media such as CD-ROM; electrical storage media such as RAM and ROM; and hybrids of these categories such as magnetic/optical storage media. A skilled artisan can readily appreciate how any of the presently known computer readable mediums can be used to create a manufacture comprising computer readable medium having recorded thereon a nucleotide sequence of the present invention.

[0132] As used herein, "recorded" refers to a process for storing information on computer readable medium. A skilled artisan can readily adopt any of the presently know methods for recording information on computer readable medium to generate manufactures comprising the nucleotide sequence information of the present invention. A variety of data storage structures are available to a skilled artisan for creating a computer readable medium having recorded thereon a nucleotide sequence of the present invention. The choice of the data storage structure will generally be based on the means chosen to access the stored information. In addition, a variety of data processor programs and formats can be used to store the nucleotide sequence information of the present invention on computer readable medium. The sequence information can be represented in a word processing text file, formatted in commercially-available software such as WordPerfect and MicroSoft Word, or represented in the form of an ASCII file, stored in a database application, such as DB2, Sybase, Oracle, or the like. A skilled artisan can readily adapt any number of dataprocessor structuring formats (e.g. text file or database) in order to obtain computer readable medium having recorded thereon the nucleotide sequence information of the present invention.

[0133] By providing the nucleotide sequence of SEQ ID NOs: 1 through 142, representative fragments thereof, or nucleotide sequences at least 99.9% identical to SEQ ID NOs: 1 through 142, in computer readable form, a skilled artisan can routinely access the sequence information for a variety of purposes. Computer software is publicly available which allows a skilled artisan to access sequence information provided in a computer readable medium. The examples which follow demonstrate how software which implements the BLAST (Altschul *et al.*, *J. Mol. Biol. 215*:403-410 (1990)) and BLAZE (Brutlag *et al.*, *Comp. Chem. 17*:203-207 (1993)) search algorithms on a Sybase system

can be used to identify open reading frames (ORFs) within the *E. coli* J96 PAI subgenome that contain homology to ORFs or proteins from other organisms. Such ORFs are protein-encoding fragments within the *E. coli* J96 PAI subgenome and are useful in producing commercially important proteins such as enzymes used in modifying surface O-antigens of bacteria. A comprehensive list of ORFs encoding commercially important *E. coli* J96 PAI proteins is provided in Tables 1 through 6.

[0134] The present invention provides a DNA sequence - gene database of pathogenicity islands (PAIs) for *E. coli* involved in infectious diseases. This database is useful for identifying and characterizing the basic functions of new virulence genes for *E. coli* involved in uropathogenic and extraintestinal diseases. The database provides a number of novel open reading frames that can be selected for further study as described herein.

[0135] Selectable insertion mutations in plasmid subclones encoding PAI genes with potentially significant phenotypes for *E. coli* uropathogenesis and sepsis can be isolated. The mutations are then crossed back into wild type, uropathogenic *E. coli* by homologous recombination to create wild-type strains specifically altered in the targeted gene. The significance of the genes to *E. coli* pathogenesis is assessed by *in vitro* assays and *in vivo* murine models of sepsis/peritonitis and ascending urinary tract infection.

[0136] New virulence genes and PAI sites in uropathogenic *E. coli* may be identified by the transposon signature-tagged mutagenesis system and negative selection of *E. coli* mutants avirulent in murine models of ascending urinary tract infection or peritonitis.

[0137] Epidemiological investigations of new virulence genes and PAIs may be used to test for their occurrence in the genomes of other pathogenic and opportunistic members of the Enterobacteriaceae.

[0138] One can choose from the ORFs included in SEQ ID NOs: 1 through 142, using Tables 1 through 6 as a useful guidepost for selecting, as candidates for targeted mutagenesis, a limited number of candidate genes within the PAIs based on their homology to virulence, export or regulation genes in other pathogens. For the large number of apparent genes within the PAIs that do not share sequence similarity to any entries in the database, the transposon signature-tagged mutagenesis method developed by David Holden's laboratory can be employed as an independent means of virulence gene identification.

[0139] Allelic knock-outs are constructed using different *pir*-dependent suicide vectors (Swihart, K.A. and R.A. Welch, *Infect. Immun. 58*:1853-1869 (1990)). In addition, two different animal model systems can be employed for assessment of pathogenic determinants. The initial identification of *E. coli* hemolysin as a virulence factor came

from the construction of isogenic *E. coli* strains that were tested in a rat model of intra-abdominal sepsis (Welch, R.A. *et al., Nature (London) 294*:665-667 (1981)). The ascending UTI (Urinary Tract Infection) mouse model was also successfully performed with allelic knock-outs of the *hpmA* hemolysin of *Proteus mirabilis* (Swihart, K.A. and R.A. Welch, *Infect. Immun. 58*:1853-1869 (1990)).

[0140] The present invention further provides systems, particularly computer-based systems, which contain the sequence information described herein. Such systems are designed to identify commercially important fragments of the *E. coli* J96 PAI subgenome. As used herein, "a computer-based system" refers to the hardware means, software means, and data storage means used to analyze the nucleotide sequence information of the present invention. The minimum hardware means of the computer-based systems of the present invention comprises a central processing unit (CPU), input means, output means, and data storage means. A skilled artisan can readily appreciate that any one of the currently available computer-based system are suitable for use in the present invention.

[0141] As indicated above, the computer-based systems of the present invention comprise a data storage means having stored therein a nucleotide sequence of the present invention and the necessary hardware means and software means for supporting and implementing a search means. As used herein, "data storage means" refers to memory that can store nucleotide sequence information of the present invention, or a memory access means which can access manufactures having recorded thereon the nucleotide sequence information of the present invention. As used herein, "search means" refers to one or more programs which are implemented on the computer-based system to compare a target sequence or target structural motif with the sequence information stored within the data storage means. Search means are used to identify fragments or regions of the *E. coli* genome that match a particular target sequence or target motif. A variety of known algorithms are disclosed publicly and a variety of commercially available software for conducting search means are available and can be used in the computer-based systems of the present invention. Examples of such software include, but are not limited to, MacPattern (EMBL), BLASTN and BLASTX (NCBIA). A skilled artisan can readily recognize that any one of the available algorithms or implementing software packages for conducting homology searches can be adapted for use in the present computer-based systems.

[0142] As used herein, a "target sequence" can be any DNA or amino acid sequence of six or more nucleotides or two or more amino acids. A skilled artisan can readily recognize that the longer a target sequence is, the less likely a target sequence will be present as a random occurrence in the database. The most preferred sequence length of a

target sequence is from about 10 to 100 amino acids or from about 30 to 300 nucleotide residues. However, it is well recognized that during searches for commercially important fragments of the *E. coli* J96 PAI subgenome, such as sequence fragments involved in gene expression and protein processing, may be of shorter length.

[0143] As used herein, "a target structural motif," or "target motif," refers to any rationally selected sequence or combination of sequences in which the sequence(s) are chosen based on a three-dimensional configuration which is formed upon the folding of the target motif. There are a variety of target motifs known in the art. Protein target motifs include, but are not limited to, enzymic active sites and signal sequences. Nucleic acid target motifs include, but are not limited to, promoter sequences, hairpin structures and inducible expression elements (protein binding sequences).

[0144] Thus, the present invention further provides an input means for receiving a target sequence, a data storage means for storing the target sequence and the homologous *E. coli* J96 PAI sequence identified using a search means as described above, and an output means for outputting the identified homologous *E. coli* J96 PAI sequence. A variety of structural formats for the input and output means can be used to input and output information in the computer-based systems of the present invention. A preferred format for an output means ranks fragments of the *E. coli* J96 PAI subgenome possessing varying degrees of homology to the target sequence or target motif. Such presentation provides a skilled artisan with a ranking of sequences which contain various amounts of the target sequence or target motif and identifies the degree of homology contained in the identified fragment.

[0145] A variety of comparing means can be used to compare a target sequence or target motif with the data storage means to identify sequence fragments of the *E. coli* J96 PAI subgenomes. For example, implementing software which implement the BLAST and BLAZE algorithms (Altschul *et al.*, *J. Mol. Biol. 215*:403-410 (1990)) can be used to identify open reading frames within the *E. coli* J96 PAI subgenome A skilled artisan can readily recognize that any one of the publicly available homology search programs can be used as the search means for the computer-based systems of the present invention.

[0146] One application of this embodiment is provided in Figure 2. Figure 2 provides a block diagram of a computer system 102 that can be used to implement the present invention. The computer system 102 includes a processor 106 connected to a bus 104. Also connected to the bus 104 are a main memory 108 (preferably implemented as random access memory, RAM) and a variety of secondary storage devices 110, such as a hard drive 112 and a removable medium storage device 114. The removable medium storage device 114 may represent, for example, a floppy disk drive, a CD-ROM drive, a magnetic

tape drive, etc. A removable storage medium 116 (such as a floppy disk, a compact disk, a magnetic tape, etc.) containing control logic and/or data recorded therein may be inserted into the removable medium storage device 114. The computer system 102 includes appropriate software for reading the control logic and/or the data from the removable medium storage device 114 once inserted in the removable medium storage device 114.

[0147] A nucleotide sequence of the present invention may be stored in a well known manner in the main memory 108, any of the secondary storage devices 110, and/or a removable storage medium 116. Software for accessing and processing the genomic sequence (such as search tools, comparing tools, etc.) reside in main memory 108 during execution.

[0148] Having generally described the invention, the same will be more readily understood by reference to the following examples, which are provided by way of illustration and are not intended as limiting.

Experimental

Example 1: High Through-put Sequencing of Cosmid Clones Covering PAI IV and PAI V in E. coli J96

[0149] The complete DNA sequence of the pathogenicity islands, PAI IV and PAI V (respectively >170 kb and ~110 kb), from uropathogenic *E. coli* strain, J96 (04:K6) was determined using a strategy, cloning and sequencing method, data collection and assembly software essentially identical to those used by the TIGR group for determining the sequence of the *Haemophilus influenzae* genome (Fleischmann, R.D., *et al.*, *Science 269*:496 (1995)). The sequences were then used for DNA and protein sequence similarity searches of the databases as described in Fleischmann, *Id.*

[0150] The analysis of the genetic information found within the PAIs of *E. coli* J96 was facilitated by the use of overlapping cosmid clones possessing these unique segments of DNA. These cosmid clones were previously constructed and mapped (as further described below) as an overlapping set in the laboratory of Dr. Doug Berg (Washington University). A gap exists between the left portion of cosmid 2 and the end of the PAI IV that would represent the *phe*V junction to the *E. coli* K-12 genome.

[0151] Uropathogenic strain *E. coli* J96 (O4:K6) was used as a source of chromosomal DNA for construction of a cosmid library. *E. coli* K-12 DH5Â and DH12 (Gibco/BRL, Gaithersburg, Md.) were used as hosts for maintaining cosmid and plasmid clones. The cosmid library of *E coli* J96 DNA was constructed essentially as described by Bukanow & Berg (*Mol. Microbiol 11*:509-523 (1994)). DNA was digested with *Sau*3AI under conditions that generated fragments with an average size of 40 to 50 kb and

electrophoresed through 1% agarose gels. Fragments of 35 to 50 kb were isolated and cloned into Lorist 6 vector that had been linearized with *Bam*III and treated with bacterial alkaline phosphatase to block self-ligation. (Lorist 6 is a 5.2-kb moderate-copy-number cosmid vector with T7 and SP6 promoters close to the cloning site.) Cloned DNA was packaged in lambda phage particles *in vitro* by using a commercial kit (Amersham, Arlington Heights, IL) and cosmid-containing phage particles were used to transduce *E. coli* DH5a. Transductant colonies were transferred to 150 mL of Luria-Bertani broth supplemented with kanamycin in 96-well microtiter plates and grown overnight at 37°C with shaking. Two sets of clones, one for each PAI were ultimately assembled, as previously described (Swenson *et al., Infection and Immunity 64*:3736-3743 (1996)), fully incorporated by reference herein).

[0152] The two sets of clones contain eleven sub-clones that were employed in the sequencing method described below. One set of four overlapping cosmid clones covers the *prs*-containing PAI V, ATCC Deposit No. 97727, deposited September 23, 1996. A second set of seven subclones covers much of the *pap*-containing PAI V, ATCC Deposit No. 97726, deposited September 23, 1996. See Figure 1.

[0153] A high throughput, random sequencing method (Fleischmann *et al., Science 269*:496 (1995); Fraser *et al., Science 270*:397 (1995)) was used to obtain the sequences for 142 (contigs) fragments of *E. coli* J96 PAIs. All clones were sequenced from both ends to aid in the eventual ordering of contigs during the sequence assembly process. Briefly, random libraries of ~ 2 kb clones covering the two J96 PAIs were constructed, ~ 2,800 clones were subjected to automated sequencing (~ 450 nt/clone) and preliminary assemblies of the sequences accomplished which result in 142 contigs for each of the two PAIs that total 95 and 135 kb respectively. The estimated sizes of the PAI IV and PAI V based on the overlapping cosmid clones are 1.7×10^5 and 1.1×10^5 bp respectively. The 142 sequences were assembled by means of the TIGR Assembler (Fleischmann *et al.*; Fraser *et al.*); Sutton *et al., Genome Sci. Tech. 1*:9 (1995)). Sequence and physical gaps were closed using a combination of strategies (Fleischmann *et al.*; Fraser *et al.*). Presently the average depth of sequencing for each base assembled in the contigs is 6-fold. The tentative identity of many genes based on sequence homology is covered in Tables 1, 3, 5 and 6.

[0154] Open reading frames (ORFs) and predicted protein-coding regions were identified as described (Fleischmann *et al.*; Fraser *et al.*) with some modification. In particular, the statistical prediction of uropathogenic *E. coli* J96 pathogenicity island genes was performed with GeneMark (Borodovsky, M. & McIninch, *J. Comput. Chem. 17*:123 (1993)). Regular GeneMark uses nonhomogeneous Markov models derived from a

training set of coding sequences and ordinary Markov models derived from a training set of noncoding sequences. The ORFs in Tables 1-6 were identified by GeneMark using a second-order Markov model trained from known *E. coli* coding regions and known *E. coli* non-coding regions. Among the important genes that are implicated in the virulence of *E. coli* J96 PAIs are adhesins, excretion pathway proteins, proteins that participate in alterations of the O-antigen in the PAIs, cytotoxins, and two-component (membrane sensor/DNA binding) proteins.

I. Adhesins.

[0155] It is believed that the principal adhesin determinants involved in uropathogenicity that are present within PAIs of uropathogenic *E. coli* are the pili encoded by the *pap*-related operons (Hultgren *et al.*, *Infect. Immun. 50*:370-377 (1993), Stromberg *et al.*, *EMBO J 9*:2001-2010 (1990), High *et al.*, *Infect. Immun. 56*:513-517 (1988)) and the distantly related afimbrial adhesins (Labigne-Roussel *et al.*, *Infect. Immun. 46*:251-259 (1988)). The presence of two of these (*pap*, and *prs*) has been confirmed. In addition potential genes for five other adhesins including *sla* (described above), AIDA-I (diffuse adherence-DEAC), *hra* (heat resistant hemagglutinin-ETEC), *fha* (filamentous hemagglutinin- *Bordetella pertussis*) and the arg-gingipain proteinase of *Porphyromonas gingivalis* have been found.

II. Type II exoprotein secretion pathway.

[0156] Highly significant statistics support the presence of multiple genes involved in the type II exoprotein pathway. Curiously, perhaps two different determinants appear to be present in PAI IV where one set of genes has the highest sequence similarity to *eps*-like genes (*Vibrio cholerae* Ctx export) and the other has greatest similarity to *exe* genes (*Aeromonas hydophilia* aerolysin and protease export). At present, the assembly of contigs involving these potential genes is incomplete. Thus, it is uncertain if two separate and complete determinants are present. However, it is clear that these genes are newly discovered and novel to pathogenic *E. coli* because the derived sequences do not have either the *bfp* or *hop* genes as the highest matches. The gene products that are the target of the type II export pathway are not evident at this time.

[0157] Within PAI IV there are sequences which suggest genes very similar to *secD* and *secF*. These two linked genes encode homologous products that are localized to the inner membrane and are hypothesized to play a late role in the translocation of leader-peptide containing proteins across the inner membrane of gram-negative bacteria. In addition, in each PAI, sequences are found that are reminiscent of the heat-shock

htrA/degA gene that encodes a piroplasmic protease. They may perform endochaperone-like function as Pugsley *et al.* have hypothesized for different exoprotein pathways.

III. O-antigen/capsule/carbohydrate modification (Nod genes).

[0158] J96 has the O4. The O-antigen portion of lipopolysaccharide is encoded by *rfb* genes that are located at 45 min. on the *E. coli* chromosome. We have found in both PAIs a cumulative total of five possible *rfb*-like genes which could participate alterations of the O-antigen in the PAIs. Overall these data suggest that PAIs provide the genetic potential for greater change of the cell surface for uropathogenic *E. coli* strains than what was previously known.

[0159] The apparent capsule type for strain J96 is a non-sialic acid K6-type. Sequence similarity "hits" were made in PAI IV region to two region-1 capsule genes, *kpsS* and *kpsE* involved in the stabilization of polysaccharide synthesis and polysaccharide export across the inner membrane. This is not altogether surprising based on the genetic mapping of the *kps* locus to *serA* at 63 minutes on the genome of the K1 capsular type of *E. coli*. This suggests that these *kps*-like genes either are participating in the K6-biosynthesis or perhaps are involved in complex carbohydrate export for other purposes.

[0160] An intriguing discovery are the hits made on genes involved in bacteria-plant interactions by *Rhizobium, Bradyrhizobium* and *Agrobacterium*. Four potential genes identified thus far share significant sequence similarity to genes encoding products that modify lipo-oligosaccharides that influence nodule morphogenesis on legume roots. These are: ORF140, carbamyl phosphate synthetase; modulation protein 1265; phosphate-regulatory protein; and an ORF at a plant-inducible locus in *Agrobacterium*. To date there are no descriptions in the literature of such gene products being utilized by human or animal bacterial pathogens for the purposes of modification or secretion of extracellular carbohydrate. However, the sequence similarity to the capsular region-2 genes and to lipooligosaccharide biosynthetic genes in *Rhizobium* spp has been recently noted by Petit (1995).

IV. Cytotoxins.

[0161] Besides the previously known hemolysin and CNF toxins in the PAIs, in each PAI sequences similar to the *shlBA* operon (cosmid 5 and 12) were found for a cytolytic toxin from *Serratia marcescens* and *Proteus mirabilis*. Ironically, the *P. mirabilis* hemolysin (HpmA) member of this family of toxins was discovered by Uphoff and Welch (1990), but not thought to exist in other members of the Enterobacteriaceae (Swihart (1990)). A *shlB*-like transporter does also appear to be involved in the export of the

filamentous hemagglutinin of *Bordetella pertussis* which was described above and a cell surface adhesin of *Haemophilus influenzae.* It has been demonstrated that cosmid #5 of *E. coli* J96 encodes an extracellular protein that is ~1 80 kDa and cross-reactive to polyclonal antisera to the *P. mirabilis* HpmA hemolysin. Thus, there is evidence suggesting there is new member of this family of proteins in extraintestinal *E. coli* isolates. In addition, there is also a hit on the FhaC hemolysin-like gene within the PAI V although its statistical significance for the sequence thus far available is only 0.0043.

V. Regulators.

[0162] A common regulatory motif in bacteria are the two-component (membrane sensor/DNA binding) proteins. In numerous instances in pathogenic bacteria, external signals in the environment cause membrane-bound protein kinases to phosphorylate a cytoplasmic protein which in turn acts as either a negative or positive effector of transcription of large sets of operons. On cosmid 11 representing PAI V were found, in two different *Pst*I clones, sequences for two-component regulators (similar probabilities for OmpR/ AIGB and separately RcsC, probabilities at the 10^{-22} level).

[0163] In addition, the phosphoglycerate transport system (*pgtA*, *pgtC*, and *pgtP*) including the *pgtB* regulator is present in PAI IV. This transport system which was originally described in *S. typhimurium* is not appreciated as a component of any pathogenic *E. coli* genome. The operon had been previously mapped at 49 minutes near or within one of the *S. typhimurium* chromosome specific-loops not present in the K-12 genome. It should be noted that the *E. coli* K-12 glpT gene product is similar to pgtP gene product (37% identity), but the *E. coli* J96 genes are clearly homologs to the *pgt* genes and their linkage within the middle of PAI IV element (cosmid #4) is suspicious.

VI. Mobile genetic elements.

[0164] There are numerous sequences that share similarity to genes found on insertion elements, plasmids and phages. The temperate bacteriophage P4 inserts within tRNA loci in the *E. coli* chromosome. The hypothesis was made that PAIs are the result of bacteriophage P4-virulence gene recombination events (Blum *et al., Infect. Immun. 62*:606-614 (1994). Data supporting this hypothesis was found during our sequencing with the identification of P4-like sequences in each of the PAIs (cosmids 7 and 9). This is a very important preliminary result which supports the hypothesis that PAIs can be identified by common sequence or genetic elements. However, there are indications that multiple mobile genetic elements involved in the evolution of the J96 PAIs. Conjugal plasmid-related sequences may also be present at two different locations (F factor and Rl

plasmid). Sequences for multiple transposable elements are present that are likely to have originated from different bacterial genera (Tnl000, IS630, IS911, IS100, IS21, IS 1203, IS5376 (*B. stearothermophflus*) and RHS). Of particular interest is IS100, which was originally identified in *Yersinia pestis* (Fetherston *et al., Mol. Microbiol. 6*:2693-2704 (1992)). The presence of IS106 is significant because it has been associated with the termini of a large chromosomal element encoding pigmentation and some aspect of virulence in *Y. pestis.* This element undergoes spontaneous deletions similar to the PAIs from *E. coli* 536 (Fetherston *et al., Mol. Microbiol. 6*:2693-2704 (1992)) and appears to participate in plasmid-chromosome rearrangements. This element was not previously known to be in genera outside of *Yersinia.*

[0165] The discovery of the apparent *att* site for bacteriophage P2 in the PAIs is interesting. P2 acts as a helper phage for the P4 satellite phage. The P2 *att* site is at 44 min in the K-12 genome. The significance of this hit is unknown at present, but may be explained as either a cloning artifact (some K-12 fragments in the *Pst* I library of cosmid 5) or evidence of some curious chromosomal-P4/ P2 phage history. It may indicate that the J96 PAIs are composites of multiple smaller PAIs.

Example 2: Preparation of PCR Primers and Amplification of DNA

[0166] Various fragments of the sequenced *E. coli* J96 PAIs, such as those disclosed in Tables 1 through 6 can be used, in accordance with the present invention, to prepare PCR primers. The PCR primers are preferably at least 15 bases, and more preferably at least 18 bases in length. When selecting a primer sequence, it is preferred that the primer pairs have approximately the same G/C ratio, so that melting temperatures are approximately the same. The PCR primers are useful during PCR cloning of the ORFs described herein.

Example 3: Gene expression from DNA Sequences Corresponding to ORFs

[0167] A fragment of an *E. coli* J96 PAIs (preferably, a protein-encoding sequence provided in Tables 1 through 6) is introduced into an expression vector using conventional technology (techniques to transfer cloned sequences into expression vectors that direct protein translation in mammalian, yeast, insect or bacterial expression systems are well known in the art). Commercially available vectors and expression systems are available from a variety of suppliers including Stratagene (La Jolla, California), Promega (Madison, Wisconsin), and Invitrogen (San Diego, California). If desired, to enhance expression and facilitate proper protein folding, the codon context and codon pairing of the sequence may be optimized for the particular expression organism, as explained by Hatfield *et al.*, U.S. Pat. No. 5,082,767, which is hereby incorporated by reference.

[0168] The following is provided as one exemplary method to generate polypeptide(s) from a cloned ORF of an *E. coli* J96 PAI whose sequence is provided in SEQ ID NOs: 1 through 142. A poly A sequence can be added to the construct by, for example, splicing out the poly A sequence from pSG5 (Stratagene) using BglI and SalI restriction endonuclease enzymes and incorporating it into the mammalian expression vector pXT1 (Stratagene) for use in eukaryotic expression systems. pXT1 contains the LTRs and a portion of the gag gene from Moloney Murine Leukemia Virus. The position of the LTRs in the construct allow efficient stable transfection. The vector includes the Herpes Simplex thymidine kinase promoter and the selectable neomycin gene. The *E. coli* J96 PAI DNA is obtained by PCR from the bacterial vector using oligonucleotide primers complementary to the *E. coli* J96 PAI DNA and containing restriction endonuclease sequences for PstI incorporated into the 5 primer and BglII at the 5 end of the corresponding *E. coli* J96 PAI DNA 3 primer, taking care to ensure that the *E. coli* J96 PAI DNA is positioned such that its followed with the poly A sequence. The purified fragment obtained from the resulting PCR reaction is digested with PstI, blunt ended with an exonuclease, digested with BglII, purified and ligated to pXT1, now containing a poly A sequence and digested BglII.

[0169] The ligated product is transfected into mouse NIH 3T3 cells using Lipofectin (Life Technologies, Inc., Grand Island, New York) under conditions outlined in the product specification. Positive transfectants are selected after growing the transfected cells in 600 ug/ml G418 (Sigma, St. Louis, Missouri). The protein is preferably released into the supernatant. However if the protein has membrane binding domains, the protein may additionally be retained within the cell or expression may be restricted to the cell surface.

[0170] Since it may be necessary to purify and locate the transfected product, synthetic 15-mer peptides synthesized from the predicted *E. coli* J96 PAI DNA sequence are injected into mice to generate antibody to the polypeptide encoded by the *E. coli* J96 PAI DNA.

[0171] If antibody production is not possible, the *E. coli* J96 PAI DNA sequence is additionally incorporated into eukaryotic expression vectors and expressed as a chimeric with, for example, ß-globin. Antibody to ß-globin is used to purify the chimeric. Corresponding protease cleavage sites engineered between the ß-globin gene and the *E. coli* J96 PAI DNA are then used to separate the two polypeptide fragments from one another after translation. One useful expression vector for generating ß-globin chimerics is pSG5 (Stratagene). This vector encodes rabbit ß-globin. Intron II of the rabbit ß-globin gene facilitates splicing of the expressed transcript, and the polyadenylation signal

incorporated into the construct increases the level of expression. These techniques as described are well known to those skilled in the art of molecular biology. Standard methods are available from the technical assistance representatives from Stratagene, Life Technologies, Inc., or Promega. Polypeptides may additionally be produced from either construct using *in vitro* translation systems such as In vitro Express^{TM} Translation Kit (Stratagene).

Example 4: E. coli Expression of an E. coli J96 PAI ORF and protein purification

[0172] An *E. coli* J96 PAI ORF described in Tables 1 through 6 is selected and amplified using PCR oligonucleotide primers designed from the nucleotide sequences flanking the selected ORF and/or from portions of the ORF s NH - or COOH-terminus. Additional nucleotides containing restriction sites to facilitate cloning are added to the 5' and 3' sequences, respectively.

[0173] The restriction sites are selected to be convenient to restriction sites in the bacterial expression vector pQE60. The bacterial expression vector pQE60 is used for bacterial expression in this example. (QIAGEN, Inc., 9259 Eton Avenue, Chatsworth, CA, 91311). pQE60 encodes ampicillin antibiotic resistance ("Ampr") and contains a bacterial origin of replication ("ori"), an IPTG inducible promoter, a ribosome binding site ("RBS"), six codons encoding histidine residues that allow affinity purification using nickel-nitrilo-tri-acetic acid ("Ni-NTA") affinity resin sold by QIAGEN, Inc., *supra,* and suitable single restriction enzyme cleavage sites. These elements are arranged such that a DNA fragment encoding a polypeptide may be inserted in such as way as to produce that polypeptide with the six His residues (i.e., a "6 X His tag") covalently linked to the carboxyl terminus of that polypeptide.

[0174] The DNA sequence encoding the desired portion of an *E. coli* J96 PAI is amplified from the deposited cDNA clone using PCR oligonucleotide primers which anneal to the amino terminal sequences of the desired portion of the *E. coli* protein and to sequences in the deposited construct 3' to the cDNA coding sequence. Additional nucleotides containing restriction sites to facilitate cloning in the pQE60 vector are added to the 5' and 3' sequences, respectively.

[0175] The amplified *E. coli* J96 PAI DNA fragments and the vector pQE60 are digested with one or more appropriate restriction enzymes, such as SalI and XbaI, and the digested DNAs are then ligated together. Insertion of the *E. coli* J96 PAI DNA into the restricted pQE60 vector places the *E. coli* J96 PAI protein coding region, including its associated stop codon, downstream from the IPTG-inducible promoter and in-frame with

an initiating AUG. The associated stop codon prevents translation of the six histidine codons downstream of the insertion point.

[0176] The ligation mixture is transformed into competent *E. coli* cells using standard procedures such as those described in Sambrook *et al., Molecular Cloning: a Laboratory Manual, 2nd Ed.;* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY (1989). *E. coli* strain Ml5/rep4, containing multiple copies of the plasmid pREP4, which expresses the lac repressor and confers kanamycin resistance ("Kanr"), is used in carrying out the illustrative example described herein. This strain, which is only one of many that are suitable for expressing an *E. coli* J96 PAI protein, is available commercially from QIAGEN, Inc., *supra.* Transformants are identified by their ability to grow on LB plates in the presence of ampicillin and kanamycin. Plasmid DNA is isolated from resistant colonies and the identity of the cloned DNA confirmed by restriction analysis, PCR and DNA sequencing.

[0177] Clones containing the desired constructs are grown overnight ("O/N") in liquid culture in LB media supplemented with both ampicillin (100 μg/ml) and kanamycin (25 μg/ml). The O/N culture is used to inoculate a large culture, at a dilution of approximately 1:25 to 1:250. The cells are grown to an optical density at 600 nm ("OD600") of between 0.4 and 0.6. isopropyl-β-D-thiogalactopyranoside ("IPTG") is then added to a final concentration of 1 mM to induce transcription from the lac repressor sensitive promoter, by inactivating the laci repressor. Cells subsequently are incubated further for 3 to 4 hours. Cells then are harvested by centrifugation.

[0178] The cells are then stirred for 3-4 hours at 4°C in 6M guanidine-HCl, pH8. The cell debris is removed by centrifugation, and the supernatant containing the *E. coli* J96 PAI protein is dialyzed against 50 mM Na-acetate buffer pH6, supplemented with 200 mM NaCl. Alternatively, the protein can be successfully refolded by dialyzing it against 500 mM NaCl, 20% glycerol, 25 mM Tris/HCl pH7.4, containing protease inhibitors. After renaturation the protein can be purified by ion exchange, hydrophobic interaction and size exclusion chromatography. Alternatively, an affinity chromatography step such as an antibody column can be used to obtain pure *E. coli* J96 PAI protein. The purified protein is stored at 4°C or frozen at –80°C.

Example 5: Cloning and Expression of an E. coli J96 PAI protein in a Baculovirus Expression System

[0179] An *E. coli* J96 PAI ORF described in Tables 1 through 6 is selected and amplified as above. The plasmid is digested with appropriate restriction enzymes and optionally, can be dephosphorylated using calf intestinal phosphatase, using routine

procedures known in the art. The DNA is then isolated from a 1% agarose gel using a commercially available kit ("Geneclean" BIO 101 Inc., La Jolla, Ca.). This vector DNA is designated herein "V1".

[0180] Fragment F1 and the dephosphorylated plasmid V1 are ligated together with T4 DNA ligase. *E. coli* HB101 or other suitable *E. coli* hosts such as XL-1 Blue (Stratagene Cloning Systems, La Jolla, CA) cells are transformed with the ligation mixture and spread on culture plates. Bacteria are identified that contain the plasmid with the *E. coli* J96 PAI gene by digesting DNA from individual colonies using appropriate restriction enzymes and then analyzing the digestion product by gel electrophoresis. The sequence of the cloned fragment is confirmed by DNA sequencing. This plasmid is designated herein pBac *E. coli J96*.

[0181] Five μg of the plasmid pBac *E. coli* J96 is co-transfected with 1.0 μg of a commercially available linearized baculovirus DNA ("BaculoGold baculovirus DNA", Pharmingen, San Diego, CA.), using the lipofection method described by Felgner *et al.*, *Proc. Natl. Acad. Sci. USA 84*:7413-7417 (1987). 1 μg of BaculoGold virus DNA and 5 μg of the plasmid pBac *E. coli* J96 are mixed in a sterile well of a microliter plate containing 50 μl of serum-free Grace's medium (Life Technologies Inc., Gaithersburg, MD). Afterwards, 10 μl Lipofectin plus 90 μl Grace's medium are added, mixed and incubated for 15 minutes at room temperature. Then the transfection mixture is added drop-wise to Sf9 insect cells (ATCC CRL 1711) seeded in a 35 mm tissue culture plate with 1 ml Grace's medium without serum. The plate is rocked back and forth to mix the newly added solution. The plate is then incubated for 5 hours at 27°C. After 5 hours the transfection solution is removed from the plate and 1 ml of Grace's insect medium supplemented with 10% fetal calf serum is added. The plate is put back into an incubator and cultivation is continued at 27°C for four days.

[0182] After four days the supernatant is collected and a plaque assay is. performed, as described by Summers and Smith, *supra.* An agarose gel with "Blue Gal" (Life Technologies Inc.) is used to allow easy identification and isolation of gal-expressing clones, which produce blue-stained plaques. (A detailed description of a "plaque assay" of this type can also be found in the user's guide for insect cell culture and baculovirology distributed by Life Technologies Inc., page 9-10). After appropriate incubation, blue stained plaques are picked with the tip of a micropipettor (e.g., Eppendorf). The agar containing the recombinant viruses is then resuspended in a microcentrifuge tube containing 200 μl of Grace's medium and the suspension containing the recombinant baculovirus is used to infect Sf9 cells seeded in 35 mm dishes. Four days later the

supernatants of these culture dishes are harvested and then they are stored at 4°C. The recombinant virus is called V-*E. coli* J96.

[0183] To verify the expression of the *E. coli* gene Sf9 cells are grown in Grace's medium supplemented with 10% heat inactivated FBS. The cells are infected with the recombinant baculovirus V-*E. coli* J96 at a multiplicity of infection ("MOI") of about 2. Six hours later the medium is removed and is replaced with SF900 II medium minus methionine and cysteine (available from Life Technologies Inc.). If radiolabeled proteins are desired, 42 hours later, 5 μCi of ^{35}S-methionine and 5 μCi ^{35}S-cysteine (available from Amersham) are added. The cells are further incubated for 16 hours and then they are harvested by centrifugation. The proteins in the supernatant as well as the intracellular proteins are analyzed by SDS-PAGE followed by autoradiography (if radiolabeled). Microsequencing of the amino acid sequence of the amino terminus of purified protein may be used to determine the amino terminal sequence of the mature protein and thus the cleavage point and length of the secretary signal peptide.

Example 6: Cloning and Expression in Mammalian Cells

[0184] Most of the vectors used for the transient expression of an *E. coli* J96 PAI gene in mammalian cells should carry the SV40 origin of replication. This allows the replication of the vector to high copy numbers in cells (e.g., COS cells) which express the T antigen required for the initiation of viral DNA synthesis. Any other mammalian cell line can also be utilized for this purpose.

[0185] A typical mammalian expression vector contains the promoter element, which mediates the initiation of transcription of mRNA, the protein coding sequence, and signals required for the termination of transcription and polyadenylation of the transcript. Additional elements include enhancers, Kozak sequences and intervening sequences flanked by donor and acceptor sites for RNA splicing. Highly efficient transcription can be achieved with the early and late promoters from SV40, the long terminal repeats (LTRS) from Retroviruses, e.g., RSV, 1HTLVI, HIVI and the early promoter of the cytomegalovirus (CMV). However, cellular elements can also be used (e.g., the human actin promoter). Suitable expression vectors for use in practicing the present invention include, for example, vectors such as PSVL and PMSG (Pharmacia, Uppsala, Sweden), pRSVcat (ATCC 37152), pSV2dhfr (ATCC 37146) and pBC12MI (ATCC 67109). Mammalian host cells that could be used include, human Hela, 293, H9 and Jurkat cells, mouse NIH3T3 and C127 cells, Cos 1, Cos 7 and CV I, quail QC1-3 cells, mouse L cells and Chinese hamster ovary (CHO) cells.

[0186] Alternatively, the gene can be expressed in stable cell lines that contain the gene integrated into a chromosome. The co-transfection with a selectable marker such as dhfr, gpt, neomycin, hygromycin allows the identification and isolation of the transfected cells.

[0187] The transfected gene can also be amplified to express large amounts of the encoded protein. The DHFR (dihydrofolate reductase) marker is useful to develop cell lines that carry several hundred or even several thousand copies of the gene of interest. Another useful selection marker is the enzyme glutamine synthase (GS) (Murphy *et al., Biochem J. 227*:277-279 (199 1); Bebbington *et al., Bio/Technology 10:*169-175 (1992)). Using these markers, the mammalian cells are grown in selective medium and the cells with the highest resistance are selected. These cell lines contain the amplified gene(s) integrated into a chromosome. Chinese hamster ovary (CHO) and NSO cells are often used for the production of proteins.

[0188] The expression vectors pC1 and pC4 contain the strong promoter (LTR) of the Rous Sarcoma Virus (Cullen *et al.,* Molecular and Cellular Biology, 438447 (March, 1985)) plus a fragment of the CMV-enhancer (Boshart *et al., Cell 41*:521-530 (1985)). Multiple cloning sites, e.g., with the restriction enzyme cleavage sites BamHI, Xbal and Asp718, facilitate the cloning of the gene of interest. The vectors contain in addition the 3' intron, the polyadenylation and termination signal of the rat preproinsulin gene.

Example 6(a): Cloning and Expression in COS Cells

[0189] The expression plasmid, p *E. coli* J96HA, is made by cloning a cDNA encoding *E. coli* J96 PAI protein into the expression vector pcDNAI/Amp or pcDNAIII (which can be obtained from Invitrogen, Inc.).

[0190] The expression vector pcDNAI/amp contains: (1) an *E. coli* origin of replication effective for propagation in *E. coli* and other prokaryotic cells; (2) an ampicillin resistance gene for selection of plasmid-containing prokaryotic cells; (3) an SV40 origin of replication for propagation in eukaryotic cells; (4) a CMV promoter, a polylinker, an SV40 intron; (5) several codons encoding a hemagglutinin fragment (i.e., an "HA" tag to facilitate purification) followed by a termination codon and polyadenylation signal arranged so that a cDNA can be conveniently placed under expression control of the CMV promoter and operably linked to the SV40 intron and the polyadenylation signal by means of restriction sites in the polylinker. The HA tag corresponds to an epitope derived from the influenza hemagglutinin protein described by Wilson *et al., Cell 37:*767 (1984). The fusion of the HA tag to the target protein allows easy detection and recovery of the

recombinant protein with an antibody that recognizes the HA epitope. pcDNAIII contains, in addition, the selectable neomycin marker.

[0191] A DNA fragment encoding the *E. coli* J96 PAI protein is cloned into the polylinker region of the vector so that recombinant protein expression is directed by the CMV promoter. The plasmid construction strategy is as follows. The *E. coli* cDNA of the deposited clone is amplified using primers that contain convenient restriction sites, much as described above for construction of vectors for expression of *E. coli* J96 PAI protein in *E. coli.*

[0192] The PCR amplified DNA fragment and the vector, pcDNAI/Amp, are digested with appropriate restriction enzymes for the chosen primer sequences and then ligated. The ligation mixture is transformed into *E. coli* strain SURE (available from Stratagene Cloning Systems, La Jolla, CA 92037), and the transformed culture is plated on ampicillin media plates which then are incubated to allow growth of ampicillin resistant colonies. Plasmid DNA is isolated from resistant colonies and examined by restriction analysis or other means for the presence of the *E. coli* J96 PAI protein-encoding fragment.

[0193] For expression of recombinant *E. coli* J96 PAI protein, COS cells are transfected with an expression vector, as described above, using DEAE-DEXTRAN, as described, for instance, in Sambrook *et al., Molecular Cloning: a Laboratory Manual,* Cold Spring Laboratory Press, Cold Spring Harbor, New York (1989). Cells are incubated under conditions for expression of *E. coli* J96 PAI protein by the vector.

[0194] Expression of the *E. coli* J96 PAI - HA fusion protein is detected by radiolabeling and immunoprecipitation, using methods described in, for example Harlow *et al., Antibodies: A Laboratory Manual, 2nd Ed.;* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York (1988). To this end, two days after transfection, the cells are labeled by incubation in media containing ^{35}S-cysteine for 8 hours. The cells and the media are collected, and the cells are washed and the lysed with detergent-containing RIPA buffer: 150 mM NaCl, 1% NP-40, 0.1% SDS, 1% NP-40, 0.5% DOC, 50 mM TRIS, pH 7.5, as described by Wilson *et al.* cited above. Proteins are precipitated from the cell lysate and from the culture media using an HA-specific monoclonal antibody. The precipitated proteins then are analyzed by SDS-PAGE and autoradiography. An expression product of the expected size is seen in the cell lysate, which is not seen in negative controls.

Example 6(b): Cloning and Expression in CHO Cells

[0195] The vector pC4 is used for the expression of an *E. coli* J96 PAI protein. Plasmid pC4 is a derivative of the plasmid pSV2-dhfr (ATCC Acc. No. 37146). The

plasmid contains the mouse DHFR gene under control of the SV40 early promoter. Chinese hamster ovary- or other cells lacking dihydrofolate activity that are transfected with these plasmids can be selected by growing the cells in a selective medium (alpha minus MEM, Life Technologies, Inc.) supplemented with the chemotherapeutic agent methotrexate. The amplification of the DHFR genes in cells resistant to methotrexate (MTX) has been well documented (see, e.g., Alt, F. W. *et al.*, 1978, *J. Biol. Chem. 253*:1357-1370, Hamlin, J. L. and Ma, C. 1990, *Biochim. et Biophys. Acta, 1097*:107-143, Page, M. J. and Sydenham, M.A. 1991, *Biotechnology 9*:64-68). Cells grown in increasing concentrations of MTX develop resistance to the drug by overproducing the target enzyme, DHFR, as a result of amplification of the DHFR gene. If a second gene is linked to the DHFR gene, it is usually co-amplified and over-expressed. It is known in the art that this approach may be used to develop cell lines carrying more than 1,000 copies of the amplified gene(s). Subsequently, when the methotrexate is withdrawn, cell lines are obtained which contain the amplified gene integrated into one or more chromosome(s) of the host cell.

[0196] Plasmid pC4 contains for expressing the gene of interest the strong promoter of the long terminal repeat (LTR) of the Rouse Sarcoma Virus (Cullen, *et al.*, *Molecular and Cellular Biology,* March 1985:438-447) plus a fragment isolated from the enhancer of the immediate early gene of human cytomegalovirus (CMV) (Boshart *et al., Cell 41*:521-530 (1985)). Downstream of the promoter is BamHI restriction enzyme site that allows the integration of the gene. Behind these cloning sites the plasmid contains the 3' intron and polyadenylation site of the rat preproinsulin gene. Other high efficiency promoters can also be used for the expression, e.g., the human -actin promoter, the SV40 early or late promoters or the long terminal repeats from other retroviruses, e.g., HIV and HTLVI. Clontech's Tet-Off and Tet-On gene expression systems and similar systems can be used to express the *E. coli* protein in a regulated way in mammalian cells (Gossen, M., & Bujard, H. 1992, *Proc. Natl. Acad. Sci. USA 89*: 5547-5551). For the polyadenylation of the mRNA other signals, e.g., from the human growth hormone or globin genes can be used as well. Stable cell lines carrying a gene of interest integrated into the chromosomes can also be selected upon co-transfection with a selectable marker such as gpt, G418 or hygromycin. It is advantageous to use more than one selectable marker in the beginning, e.g., G418 plus methotrexate.

[0197] The plasmid pC4 is digested with appropriate restriction enzymes and then dephosphorylated using calf intestinal phosphates by procedures known in the art. The vector is then isolated from a 1% agarose gel.

[0198] The DNA sequence encoding the complete *E. coli* J96 PAI protein including its leader sequence is amplified using PCR oligonucleotide primers corresponding to the 5' and 3' sequences of the gene.

[0199] The amplified fragment is digested with appropriate endonucleases for the chosen primers and then purified again on a 1% agarose gel. The isolated fragment and the dephosphorylated vector are then ligated with T4 DNA ligase. *E. coli* HB101 or XL-1 Blue cells are then transformed and bacteria are identified that contain the fragment inserted into plasmid pC4 using, for instance, restriction enzyme analysis.

[0200] Chinese hamster ovary cells lacking an active DHFR gene are used for transfection. 5 µg of the expression plasmid pC4 is cotransfected with 0.5 µg of the plasmid pSVneo using lipofectin (Felgner *et al., supra).* The plasmid pSV2neo contains a dominant selectable marker, the neo gene from Tn5 encoding an enzyme that confers resistance to a group of antibiotics including G418. The cells are seeded in alpha minus MEM supplemented with 1 mg/ml G418. After 2 days, the cells are trypsinized and seeded in hybridoma cloning plates (Greiner, Germany) in alpha minus MEM supplemented with 10, 25, or 50 ng/ml of methothrexate plus 1 mg/ml G418. After about 10-14 days single clones are trypsinized and then seeded in 6-well petri dishes or 10 ml flasks using different concentrations of methotrexate (50 nM, 100 nM, 200 nM, 400 nm, 800 nM). Clones growing at the highest concentrations of methotrexate are then transferred to new 6-well plates containing even higher concentrations of methotrexate (1 µM, 2 µM, 5 µM, 10 mM, 20 mM). The same procedure is repeated until clones are obtained which grow at a concentration of 100 - 200 µM. Expression of the desired gene product is analyzed, for instance, by SDS-PAGE and Western blot or by reversed phase HPLC analysis.

Example 7: Production of an Antibody to an E. coli J96 Pathogenicity Island Protein

[0201] Substantially pure *E. coli* J96 PAI protein or polypeptide is isolated from the transfected or transformed cells described above using an art-known method. The protein can also be chemically synthesized. Concentration of protein in the final preparation is adjusted, for example, by concentration on an Amicon filter device, to the level of a few micrograms/ml. Monoclonal or polyclonal antibody to the protein can then be prepared as follows:

I. Monoclonal Antibody Production by Hybridoma Fusion

[0202] Monoclonal antibody to epitopes of any of the peptides identified and isolated as described can be prepared from murine hybridomas according to the classical method of

Kohler and Milstein, *Nature 256*:495 (1975) or modifications of the methods thereof. Briefly, a mouse is repetitively inoculated with a few micrograms of the selected protein over a period of a few weeks. The mouse is then sacrificed, and the antibody producing cells of the spleen isolated. The spleen cells are fused by means of polyethylene glycol with mouse myeloma cells, and the excess unfused cells destroyed by growth of the system on selective media comprising aminopterin (HAT media). The successfully fused cells are diluted and aliquots of the dilution placed in wells of a microtiter plate where growth of the culture is continued. Antibody-producing clones are identified by detection of antibody in the supernatant fluid of the wells by immunoassay procedures, such as ELISA, as originally described by Engvall, E., *Meth. Enzymol. 70*:419 (1980), and modified methods thereof. Selected positive clones can be expanded and their monoclonal antibody product harvested for use. Detailed procedures for monoclonal antibody production are described in Davis, L. *et al.* Basic Methods in Molecular Biology Elsevier, New York. Section 21-2 (1989).

II. Polyclonal Antibody Production by Immunization

[0203] Polyclonal antiserum containing antibodies to heterogenous epitopes of a single protein can be prepared by immunizing suitable animals with the expressed protein described above, which can be unmodified or modified to enhance immunogenicity. Effective polyclonal antibody production is affected by many factors related both to the antigen and the host species. For example, small molecules tend to be less immunogenic than other molecules and may require the use of carriers and adjuvant. Also, host animals vary in response to site of inoculations and dose, with both inadequate or excessive doses of antigen resulting in low titer antisera. Small doses (ng level) of antigen administered at multiple intradermal sites appears to be most reliable. An effective immunization protocol for rabbits can be found in Vaitukaitis, J. *et al.*, *J. Clin. Endocrinol. Metab. 33*:988-991 (1971).

[0204] Booster injections can be given at regular intervals, and antiserum harvested when antibody titer thereof, as determined semi-quantitatively, for

[0205] example, by double immunodiffusion in agar against known concentrations of the antigen, begins to fall (*See* Ouchterlony, O. *et al.*, Chap. 19 in: *Handbook of Experimental Immunology*, Wier, D., ed, Blackwell (1973)). Plateau concentration of antibody is usually in the range of 0.1 to 0.2 mg/ml of serum (about 12 μM). Affinity of the antisera for the antigen is determined by preparing competitive binding curves, as described, for example, by Fisher, D., Chap. 42 in: *Manual of Clinical Immunology*, 2nd ed., Rose and Friedman, (eds.), Amer. Soc. For Microbio., Washington, D.C. (1980).

[0206] Antibody preparations prepared according to either protocol are useful in quantitative immunoassays which determine concentrations of antigen-bearing substances in biological samples; they are also used semi-quantitatively or qualitatively to identify the presence of antigen in a biological sample.

[0207] While the present invention has been described in some detail for purposes of clarity and understanding, one skilled in the art will appreciate that various changes in form and detail can be made without departing from the true scope of the invention.

[0208] All patents, patent applications and publications recited herein are hereby incorporated by reference.

TABLE 1 (PAI IV)

Putative coding regions of novel *E. coli* PAI IV proteins similar to known proteins

Contig ID	ORF ID	Start (nt)	Stop (nt)	match acession	match gene name	% sim	% ident	length (nt)
65	2	1902	1042	gi|1655838	ORFB; putative transposase (Yersinia pestis)	100	100	861
65	3	2096	1821	gi|467612	ORF1 (Yersinia pestis)	100	100	276
63	11	7856	9238	gi|154262	transporter protein pgtP (Salmonella typhimurium)	98	93	1383
65	4	2889	1915	gi|1655837	ORFA; putative transposase (Yersinia pestis)	97	96	975
138	1	2	172	gi|1208992	unknown (Escherichia coli)	97	78	171
64	6	4075	4338	gi|1143207	Description: IS630 insertion element; ORF5 protein; Method: conceptual translation supplied by author (Shigella sonnei)	92	92	264
67	1	1	273	gi|809648	ExeF gene product (Aeromonas hydrophila)	92	71	273
73	4	3029	2511	gi|799234	glucose-1-phosphate thymidylyltransferase (Escherichia coli)	92	86	519
73	5	3139	2996	gi|454900	rfbC gene product (Shigella flexneri)	92	92	144
64	5	3741	4088	gi|47542	ORF (343 AA) (Shigella sonnei)	91	85	348
73	3	2613	2242	gi|46985	glucose-1-phosphate thymidylyltransferase (Salmonella enterica)	91	82	372
90	1	1	366	gi|38826	ExeE gene product (Aeromonas hydrophila)	91	77	366
91	2	604	248	gi|609625	putative (Vibrio cholerae)	91	67	357
63	9	6301	5234	gi|858753	regulatory protein pgtB (Salmonella typhimurium)	89	84	1068
73	2	2179	1811	gi|294899	dTDP-6-deoxy-L-mannose-dehydrogenase (Shigella flexneri)	89	84	369
90	2	201	689	gi|38826	ExeE gene product (Aeromonas hydrophila)	89	80	489
95	2	1519	413	gi|581654	dTDP-glucose 4,6-dehydratase (Salmonella enterica)	88	81	1107
96	1	729	457	pir|S43483|S434	Orf104 homolog - Escherichia coli	88	72	273
63	6	4281	3019	gi|154255	phosphoglycerate transport system activator protein (Salmonella typhimurium)	87	79	1263
67	2	251	745	gi|609628	putative (Vibrio cholerae)	87	72	495
82	12	5254	4406	gi|1208992	unknown (Escherichia coli)	87	74	849
60	1	693	4	gi|609625	putative (Vibrio cholerae)	86	57	690
95	1	428	3	gi|508238	dTDP-6-deoxy-L-mannose-dehydrogenase (Escherichia coli)	85	74	426
64	7	4336	4731	gi|47542	ORF (343 AA) (Shigella sonnei)	84	81	396
80	8	2800	2582	gi|38832	ExeK gene product (Aeromonas hydrophila)	84	53	219
82	10	4380	3829	gi|1033137	ORF_o152 (Escherichia coli)	84	72	552

TABLE 1 (PAI IV)(CONTINUED)

Contig ID	ORF ID	Start (nt)	Stop (nt)	match acession	match gene name	% sim	% ident	length (nt)
63	8	5399	4830	sp|P37433|PGTB_	PHOSPHOGLYCERATE TRANSPORT SYSTEM SENSOR PROTEIN PGTB (EC 2.7.3.-).	83	75	570
63	10	7572	6259	gi|154258	regulatory protein pgtC (Salmonella typhimurium)	83	78	1314
65	7	3351	3100	gi|1196999	unknown protein (Transposon Tn3411)	82	80	252
100	1	337	2	gi|41004	URF 2 (Escherichia coli)	82	64	336
138	2	109	429	gi|1033128	ORF_o273 (Escherichia coli)	80	62	321
74	4	1331	831	gi|38826	ExeE gene product (Aeromonas hydrophila)	79	62	501
63	7	4873	4256	sp|P37433|PGTB_	PHOSPHOGLYCERATE TRANSPORT SYSTEM SENSOR PROTEIN PGTB (EC 2.7.3.-).	78	72	618
70	13	5759	5529	gi|1773143	Hha protein (Escherichia coli)	78	58	231
91	3	1154	534	gi|609625	putative (Vibrio cholerae)	77	65	621
75	5	3524	3255	gi|463911	heat resistant agglutinin 1 (Escherichia coli)	76	62	270
63	1	2	667	gi|1574313	H. influenzae predicted coding region HI1472 (Haemophilus influenzae)	75	56	666
104	2	485	315	gi|530438	arabinose transport protein (Mycoplasma capricolum)	72	41	171
63	3	2180	1629	gi|622948	transposase (Escherichia coli)	71	60	552
63	12	9688	10005	sp|P39213|YI91_	INSERTION ELEMENT IS911 HYPOTHETICAL 12.7 KD PROTEIN.	71	57	318
61	3	1283	876	gi|581535	ORF140 gene product (Rhizobium sp.)	70	54	408
84	3	2361	3437	gi|1772623	HecA (Erwinia chrysanthemi)	70	60	1077
91	1	300	4	gi|295430	epsE (Vibrio cholerae)	70	49	297
74	1	541	2	gi|609627	putative (Vibrio cholerae)	69	54	540
67	4	1297	1581	gi|151469	PilD-dependent protein (Pseudomonas aeruginosa)	68	50	285
84	1	578	1741	gi|1772622	HecB (Erwinia chrysanthemi)	68	54	1164
84	2	1698	2363	gi|1772622	HecB (Erwinia chrysanthemi)	67	48	666
63	2	1734	1393	gi|1323798	transposase (Plasmid pRL1063a)	65	46	342
71	1	1134	4	gi|397405	kpsE gene product (Escherichia coli)	65	36	1131
64	2	2828	1839	gi|310632	hydrophobic membrane protein (Streptococcus gordonii)	64	38	990
74	2	861	355	gi|148436	secretory component (Erwinia chrysanthemi)	64	54	507
66	1	556	2	gi|1235662	RfbC (Myxococcus xanthus)	62	39	555
70	6	3017	2814	gi|1657478	similar to E. coli ORF_o208 (Escherichia coli)	62	41	204

TABLE 1 (PAI IV)(CONTINUED)

Contig ID	ORF ID	Start (nt)	Stop (nt)	match accession	match gene name	% sim	% ident	length (nt)
85	1	278	66	pir|A45253|A452	activator 1 37K chain - human	62	56	213
126	1	3	323	gi|1778562	hypothetical protein (Escherichia coli)	62	45	321
73	1	773	3	pir|S32879|S328	lipA protein - Neisseria meningitidis	61	46	771
96	2	796	644	gnl|PID|e276217	T03F6.f (Caenorhabditis elegans)	61	46	153
67	3	743	1312	gi|609629	putative (Vibrio cholerae)	60	43	570
70	10	4666	4292	gi|1657478	similar to E. coli ORF_o208 (Escherichia coli)	60	45	375
81	1	1	1179	gi|1591717	spore coat polysaccharide biosynthesis protein E (Methanococcus jannaschii)	60	44	1179
80	5	2563	1790	gi|609632	putative (Vibrio cholerae)	59	41	774
137	1	73	528	gi|1736670	Adhesin AIDA-I precursor. (Escherichia coli)	59	45	456
61	1	773	3	gi|1196968	unknown protein (Insertion sequence IS66)	58	41	771
63	5	2831	2178	gi|622948	transposase (Escherichia coli)	58	41	654
64	3	3568	2690	gi|1335913	unknown (Erysipelothrix rhusiopathiae)	57	36	879
64	1	1819	917	gi|153826	adhesin B (Streptococcus sanguis)	55	30	903
64	9	7008	6685	gi|152259	lcrB gene product (Rhizobium sp.)	55	42	324
70	14	6481	6753	pir|G42465|G424	hypothetical protein 88 - phage phi-R73	53	30	273
85	5	9317	1530	gi|144048	filamentous hemagglutinin (Bordetella pertussis)	52	37	7788
64	8	5063	4806	gnl|PID|e264304	F53C11.6 (Caenorhabditis elegans)	51	27	258
80	9	3411	2761	gi|149309	pulJ (Klebsiella pneumoniae)	50	40	651
88	1	98	388	gi|156087	(Brugia malayi myosin heavy chain gene, complete cds.), gene product (Brugia malayi)	50	32	291
96	3	1127	687	gi|1196964	unknown protein (Plasmid Ti)	50	38	441
89	1	981	4	gi|57633	neuronal myosin heavy chain (Rattus rattus)	48	22	978
113	1	657	199	gi|147899	extragenic suppressor (Escherichia coli)	48	25	459
118	1	654	145	pir|S27564|S275	polysaccharide translocation-related protein - Escherichia coli	48	25	510
58	2	2101	4245	gi|1235662	RfbC (Myxococcus xanthus)	47	35	2145
87	1	595	134	gi|1235662	RfbC (Myxococcus xanthus)	42	28	462
85	2	1018	515	bbs|117606	glycine-rich protein, atGRP (clone atGRP-1) (Arabidopsis thaliana, C24, Peptide Partial, 210 aa) (Arabidopsis thaliana)	36	36	504
85	3	1779	973	bbs|157676	silk fibroin heavy chain (C-terminal) (Bombyx mori-silkworms, Peptide Partial, 633 aa) (Bombyx mori)	34	29	807

TABLE 2 (PAI IV)

Putative coding regions of novel *E. coli* PAI IV proteins not similar to known proteins

Contig ID	ORF ID	Start (nt)	Stop (nt)
58	1	1176	2120
61	2	54	560
63	4	1875	2639
64	4	3911	3627
65	6	3009	3239
65	12	6027	6683
66	2	1289	978
70	2	1418	861
70	3	1886	1476
70	4	2124	1900
70	5	2795	2220
70	7	3645	3259
70	8	4078	3680
70	9	4220	4513
70	11	4950	4498
70	12	4594	4866
70	15	6805	7449
70	16	9520	10806
73	7	3247	3666
74	3	720	1301
75	1	1	165
79	1	719	354
80	6	2108	2575
80	7	2831	2469
80	10	3223	3387
80	11	3541	3362
82	8	3313	4260
82	11	4340	5218
82	13	6090	5614
84	4	3487	3281
85	4	1485	2285
85	6	8373	9320
104	1	358	2
112	1	677	105
142	1	3	143
142	2	119	328

TABLE 3 (PAI V)

Putative coding regions of novel *E. coli* PAI V proteins similar to known proteins

Contig ID	ORF ID	Start (nt)	Stop (nt)	match accession	match gene name	% sim	% ident	length (nt)
14	3	2826	3686	gi|1655838	ORFB; putative transposase [Yersinia pestis]	100	100	861
14	2	1837	2907	gi|1655837	ORFA; putative transposase [Yersinia pestis]	99	99	1071
3	9	7927	7595	gi|1657499	putative transposase for insertion sequence IS3 [Escherichia coli]	89	85	333
20	6	3462	4304	gi|1208992	unknown [Escherichia coli]	87	73	843
6	6	3541	3263	pir|S43483|S434	Orf104 homolog - Escherichia coli	81	62	279
20	3	1616	2332	gi|1033129	ORF_o233 [Escherichia coli]	80	61	717
9	1	1	681	gi|537112	ORF_o196 [Escherichia coli]	77	55	681
15	3	1899	1672	pir|S43483|S434	Orf104 homolog - Escherichia coli	75	55	228
20	9	4302	4880	gi|1552816	similar to E. coli ORF_o152 [Escherichia coli]	74	60	579
14	13	12972	15359	gi|1772623	HecA [Erwinia chrysanthemi]	70	60	2388
5	3	4112	1570	gi|1001717	regulatory components of sensory transduction system [Synechocystis sp.]	68	45	459
3	1	2572	1373	gi|849022	Lactate oxidase [Aerococcus viridans]	66	46	1200
3	8	6869	6498	gi|581535	ORF140 gene product [Rhizobium sp.]	66	45	372
6	5	3265	2951	gi|642184	F19C6.1 [Caenorhabditis elegans]	66	44	315
14	12	11775	12974	gi|1772622	HecB [Erwinia chrysanthemi]	66	50	1200
20	1	545	1450	gi|1033127	ORF_o289 [Escherichia coli]	66	45	906
57	1	696	124	gi|1772622	HecB [Erwinia chrysanthemi]	66	47	573
3	3	3320	3700	gi|431950	similar to a B.subtilis gene (GB; BACHEMEHY_5) [Clostridium pasteurianum]	65	34	381
5	7	4565	4239	sp|P39213|YI91_	INSERTION ELEMENT IS911 HYPOTHETICAL 12.7 KD PROTEIN.	65	38	327
22	2	1651	557	gi|290430	adhesin [Escherichia coli]	64	48	1095
5	4	1455	1841	gi|1575577	DNA-binding response regulator [Thermotoga maritima]	61	47	387
14	11	11161	11937	gi|1772622	HecB [Erwinia chrysanthemi]	60	39	777
14	1	930	1700	gi|1657478	similar to E. coli ORF_o208 [Escherichia coli]	58	47	771
5	6	3834	3391	gi|155032	ORF B [Plasmid pEa34]	56	36	444
3	5	6500	5982	gi|1633572	Herpesvirus saimiri ORF73 homolog [Kaposi's sarcoma-associated herpes-like virus]	54	25	519
14	7	8429	8809	gi|1196729	unknown protein [Bacteriophage P4]	54	41	381

TABLE 3 (PAI V) (CONTINUED)

Contig ID	ORF ID	Start (nt)	Stop (nt)	match accession	match gene name	% sim	% ident	length (nt)
14	14	15191	21793	gi|144048	filamentous hemagglutinin (Bordetella pertussis)	52	37	6603
14	16	21427	22671	bbs|117613	glycine-rich protein, atGRP (clone atGRP-4) (Arabidopsis thaliana, C24, Peptide Partial, 112 aa) (Arabidopsis thaliana)	52	39	1245
5	2	1004	381	gi|48518	HydC (Wolinella succinogenes)	51	34	624
5	5	1941	3311	gi|143331	alkaline phosphatase regulatory protein (Bacillus subtilis)	51	21	1371
14	4	3968	5431	gi|1033120	ORF_o469 (Escherichia coli)	51	29	1464
32	1	481	227	gi|1673731	(AE000010) Mycoplasma pneumoniae, fructose-permease IIBC component; similar to Swiss-Prot Accession Number P20966, from E. coli (Mycoplasma pneumoniae)	50	41	255
20	17	7039	7284	gi|1123054	coded for by C. elegans cDNA CEESN53F; similar to protein kinases including CDC15 in yeast (Caenorhabditis elegans)	48	28	246

TABLE 4 (PAI V)

Putative coding regions of novel *E. coli* PAI V proteins not similar to known proteins

Contig ID	ORF ID	Start (nt)	Stop (nt)
1	1	809	1165
3	2	3275	2640
3	6	6006	6425
3	7	6423	6833
4	1	3	455
5	1	501	4
6	1	2168	1749
6	2	2527	2114
6	3	2648	2331
6	4	3099	2626
14	5	7112	7699
14	6	7800	8507
14	8	9040	9624
14	10	10586	10846
14	15	21721	20921
15	1	575	826
15	2	850	1365
20	2	904	605
20	4	2330	3157
20	5	3139	3396
20	7	3812	3492
20	8	4373	3828
20	18	7282	7950
22	1	356	3
24	1	492	4

TABLE 5 (PAI IV)

Putative coding regions of novel *E. coli* PAI IV containing known *E. coli* sequences

Contig ID	ORF ID	Start (nt)	Stop (nt)	match acession	match gene name	percent ident	HSP nt length	ORF nt length
59	1	968	54	emb|X61239|ECPA	E.coli papABCDEFGHIJK genes for F13 P-pili proteins	99	790	915
59	2	1551	805	emb|Y00529|ECPA	E. coli papC gene involved in formation of pap pili	99	518	747
59	3	1742	1494	emb|Y00529|ECPA	E. coli papC gene involved in formation of pap pili	99	182	249
61	4	1975	1220	emb|X61239|ECPA	E.coli papABCDEFGHIJK genes for F13 P-pili proteins	100	69	756
63	13	10097	10480	gb|AE000133|	Escherichia coli from bases 263572 to 274477 (section 23 of 400) of the complete genome	91	216	384
65	1	886	671	gb|U06468|	Escherichia coli O111:H- insertion sequence IS1203 12.7 kDa protein and putative transposase genes, complete cds	93	164	216
65	5	3218	2868	gb|U06468|	Escherichia coli O111:H- insertion sequence IS1203 12.7 kDa protein and putative transposase genes, complete cds	85	285	351
65	8	4064	3216	gb|U06468|	Escherichia coli O111:H- insertion sequence IS1203 12.7 kDa protein and putative transposase genes, complete cds	86	145	849
65	9	4939	4337	emb|Y00976|ECHN	E.coli hns gene for DNA-binding protein H-NS (5'-region)	96	53	603
65	10	4939	5266	emb|Y00976|ECHN	E.coli hns gene for DNA-binding protein H-NS (5'-region)	98	310	348
65	11	5206	5781	gb|AE000133|	Escherichia coli from bases 263572 to 274477 (section 23 of 400) of the complete genome	89	431	576
68	1	1575	1315	emb|X61239|ECPA	E.coli papABCDEFGHIJK genes for F13 P-pili proteins	100	186	261
68	2	2468	1848	emb|X51704|ECPA	Escherichia coli papJ gene for PapJ protein	99	621	621
68	3	2232	2594	emb|X61239|ECPA	E.coli papABCDEFGHIJK genes for F13 P-pili proteins	99	363	363
68	4	3212	2466	emb|X61239|ECPA	E.coli papABCDEFGHIJK genes for F13 P-pili proteins	100	747	747
69	1	300	4	gb|M14040|	E.coli apt gene encoding adenine phosphoribosyl-transferase (APRT), complete cds	98	225	297
69	2	383	117	gb|M14040|	E.coli apt gene encoding adenine phosphoribosyl-transferase (APRT), complete cds	95	162	267
70	1	832	149	gb|U09857|	Escherichia coli 4787 o115;v165;f165 fimbrial regulatory f16521, f16528 and f1652 A genes, complete cds	89	225	684
70	17	10799	11767	gb|AE000291|	Escherichia coli , asnV, erfK, cobT, cobS, cobU, yi52_6, yi22_3, yi21_3 genes from bases 2060089 to 2072765 (section 181 of 400) of the complete genome	95	553	969
70	18	11809	11045	gb|AE000291|	Escherichia coli , asnV, erfK, cobT, cobS, cobU, yi52_6, yi22_3, yi21_3 genes from bases 2060089 to 2072765 (section 181 of 400) of the complete genome	94	595	765
70	19	12022	15222	dbj|D90838|D908	E.coli genomic DNA, Kohara clone #348(44.5-44.9 min.)	89	2667	3201

TABLE 5 (PAI IV) (CONTINUED)

Contig ID	ORF ID	Start (nt)	Stop (nt)	match acession	match gene name	percent ident	HSP nt length	ORF nt length
70	20	15316	16836	gb|AE000292|	Escherichia coli , yeeA, sbmC, yeeC, sbcB, yeeD, yeeE genes from bases 2072708 to 2083664 (section 182 of 400) of the complete genome	96	1488	1521
70	21	16722	17711	gb|AE000292|	Escherichia coli , yeeA, sbmC, yeeC, sbcB, yeeD, yeeE genes from bases 2072708 to 2083664 (section 182 of 400) of the complete genome	96	82	990
70	22	17426	16776	gb|AE000292|	Escherichia coli , yeeA, sbmC, yeeC, sbcB, yeeD, yeeE genes from bases 2072708 to 2083664 (section 182 of 400) of the complete genome	96	63	651
72	1	12	1061	gb|M10133|	E.coli (J96) hlyC, hlyA, hlyB and hlyD genes coding for chromosomal hemolysins C, A, B and D	99	1024	1050
72	2	947	1285	gb|M10133|	E.coli (J96) hlyC, hlyA, hlyB and hlyD genes coding for chromosomal hemolysins C, A, B and D	96	261	339
73	6	4437	3205	gb|AE000379|	Escherichia coli from bases 3102169 to 3112339 (section 269 of 400) of the complete genome	95	392	1233
73	8	6177	4555	gb|U28377|	Escherichia coli K-12 genome; approximately 65 to 68 minutes	90	1133	1623
73	9	6835	6128	gb|AE000380|	Escherichia coli , glcB, glcG, glcD genes from bases 3112500 to 3126189 (section 270 of 400) of the complete genome	93	703	708
75	2	1553	1059	gb|AE000498|	Escherichia coli from bases 4493507 to 4503769 (section 388 of 400) of the complete genome	90	385	495
75	3	2579	1566	gb|AE000498|	Escherichia coli from bases 4493507 to 4503769 (section 388 of 400) of the complete genome	92	464	1014
75	4	3297	2743	gb|U07174|	Escherichia coli O9:H10:K99 heat resistant agglutinin 1 gene, complete cds	81	283	555
76	1	698	3	gb|M10133|	E.coli (J96) hlyC, hlyA, hlyB and hlyD genes coding for chromosomal hemolysins C, A, B and D	99	693	696
78	1	382	59	gb|AE000360|	Escherichia coli from bases 2885166 to 2897277 (section 250 of 400) of the complete genome	99	315	324
79	2	2620	1529	gb|M10133|	E.coli (J96) hlyC, hlyA, hlyB and hlyD genes coding for chromosomal hemolysins C, A, B and D	99	1084	1092
79	3	2925	2587	gb|M10133|	E.coli (J96) hlyC, hlyA, hlyB and hlyD genes coding for chromosomal hemolysins C, A, B and D	97	322	339
79	4	3576	2923	gb|M10133|	E.coli (J96) hlyC, hlyA, hlyB and hlyD genes coding for chromosomal hemolysins C, A, B and D	99	654	654
80	1	376	83	gb|U05251|	Escherichia coli polysialic acid gene cluster region 3, promoter region	93	210	294
80	2	638	210	gb|AE000379|	Escherichia coli from bases 3102169 to 3112339 (section 269 of 400) of the complete genome	95	347	429
80	3	1246	710	gb|AE000379|	Escherichia coli from bases 3102169 to 3112339 (section 269 of 400) of the complete genome	96	388	537

TABLE 5 (PAI IV) (CONTINUED)

Contig ID	ORF ID	Start (nt)	Stop (nt)	match acession	match gene name	percent idenc	HSP nt length	ORF nt length
80	4	1796	1182	gb|AE000379|	Escherichia coli from bases 3102169 to 3112339 (section 269 of 400) of the complete genome	94	397	615
82	1	1	567	emb|X74567|ECKP	E.coli K5 antigen gene cluster region 1 kpsE, kpsD, kpsU, kpsC and kpsS genes	87	551	567
82	2	549	1157	emb|X74567|ECKP	E.coli K5 antigen gene cluster region 1 kpsE, kpsD, kpsU, kpsC and kpsS genes	88	554	609
82	3	1500	1180	gb|AE000292|	Escherichia coli , yeeA, sbmC, yeeC, sbcB, yeeD, yeeE genes from bases 2072708 to 2083664 (section 182 of 400) of the complete genome	90	62	321
82	4	2163	1519	gb|AE000292|	Escherichia coli , yeeA, sbmC, yeeC, sbcB, yeeD, yeeE genes from bases 2072708 to 2083664 (section 182 of 400) of the complete genome	89	143	645
82	5	2594	2139	gb|AE000292|	Escherichia coli , yeeA, sbmC, yeeC, sbcB, yeeD, yeeE genes from bases 2072708 to 2083664 (section 182 of 400) of the complete genome	97	456	456
82	6	3000	2605	gb|AE000292|	Escherichia coli , yeeA, sbmC, yeeC, sbcB, yeeD, yeeE genes from bases 2072708 to 2083664 (section 182 of 400) of the complete genome	98	396	396
82	7	3463	3047	gb|AE000292|	Escherichia coli , yeeA, sbmC, yeeC, sbcB, yeeD, yeeE genes from bases 2072708 to 2083664 (section 182 of 400) of the complete genome	96	283	417
82	9	3831	3337	gb|AE000292|	Escherichia coli , yeeA, sbmC, yeeC, sbcB, yeeD, yeeE genes from bases 2072708 to 2083664 (section 182 of 400) of the complete genome	96	453	495
83	1	3	311	gb|AE000151|	Escherichia coli , ybaE, cof, mdlA, mdlB, glnK, amtB, tasB, ffs genes from bases 464774 to 475868 (section 41 of 400) of the complete genome	99	207	309
83	2	176	433	gb|AE000151|	Escherichia coli , ybaE, cof, mdlA, mdlB, glnK, amtB, tasB, ffs genes from bases 464774 to 475868 (section 41 of 400) of the complete genome	100	223	258
86	1	529	2	gb|AE000379|	Escherichia coli from bases 3102169 to 3112339 (section 269 of 400) of the complete genome	93	398	528
93	1	440	3	gb|M10133|	E.coli (J96) hlyC, hlyA, hlyB and hlyD genes coding for chromosomal hemolysins C, A, B and D	95	351	438
94	1	368	72	emb|X14180|ECGL	Escherichia coli glutamine permease glnHPQ operon	100	229	297
99	1	161	586	gb|AE000379|	Escherichia coli from bases 3102169 to 3112339 (section 269 of 400) of the complete genome	98	426	426
99	2	643	476	gb|AE000379|	Escherichia coli from bases 3102169 to 3112339 (section 269 of 400) of the complete genome	99	168	168
99	3	532	1092	gb|AE000379|	Escherichia coli from bases 3102169 to 3112339 (section 269 of 400) of the complete genome	95	537	561
99	4	1094	1396	gb|AE000379|	Escherichia coli from bases 3102169 to 3112339 (section 269 of 400) of the complete genome	94	274	303

TABLE 5 (PAI IV) (CONTINUED)

Contig ID	ORF ID	Start (nt)	Stop (nt)	match acession	match gene name	percent ident	HSP nt length	ORF nt length
102	1	527	3	emb|Y00529|ECPA	E. coli papC gene involved in formation of pap pili	100	427	525
102	2	762	373	emb|Y00529|ECPA	E. coli papC gene involved in formation of pap pili	99	333	390
105	1	377	3	gb|AE000480|	Escherichia coli from bases 4277211 to 4288813 (section 370 of 400) of the complete genome	100	343	375
107	1	2	397	gb|M10133|	E.coli (J96) hlyC, hlyA, hlyB and hlyD genes coding for chromosomal hemolysins C, A, B and D	99	390	396
107	2	406	966	gb|M10133|	E.coli (J96) hlyC, hlyA, hlyB and hlyD genes coding for chromosomal hemolysins C, A, B and D	99	549	561
110	1	148	2	emb|X56175|ECSE	Escherichia coli secD and secF genes for membrane proteins involved in protein export	99	143	147
110	2	312	40	gb|M63939|	E.coli tRNA-guanine-transglycosylase (tgt) gene, complete cds	100	125	273
115	1	501	325	gb|AE000459|	Escherichia coli from bases 4013123 to 4024654 (section 349 of 400) of the complete genome	98	177	177
117	1	3	302	gb|AE000506|	Escherichia coli from bases 4584059 to 4594314 (section 396 of 400) of the complete genome	100	263	300
121	1	2	250	gb|M16202|	E.coli papH gene encoding a pilin-like protein	98	148	249
123	1	361	2	gb|AE000379|	Escherichia coli from bases 3102169 to 3112339 (section 269 of 400) of the complete genome	99	113	360
127	1	2	229	gb|AE000233|	Escherichia coli , racC, ydaD, sieB, trkG genes from bases 1415432 to 1425731 (section 123 of 400) of the complete genome	100	200	228
127	2	227	382	gb|AE000233|	Escherichia coli , racC, ydaD, sieB, trkG genes from bases 1415432 to 1425731 (section 123 of 400) of the complete genome	97	113	156
130	1	337	2	emb|X60200|ECTN	E. coli transposon Tn1000 (gamma delta) tnpR and tnpA genes for resolvase and transposase	99	335	336
131	1	510	79	gb|M30198|	E.coli recQ gene complete cds, and pldA gene, 3' end	98	304	432
131	2	743	270	gb|M30198|	E.coli recQ gene complete cds, and pldA gene, 3' end	99	314	474
133	1	1	258	gb|AE000115|	Escherichia coli , yabF, kefC, folA, apaH, apaG, ksgA, pdxA, surA, imp genes from bases 47163 to 57264 (section 5 of 400) of the complete genome	98	237	258
133	2	192	350	gb|AE000115|	Escherichia coli , yabF, kefC, folA, apaH, apaG, ksgA, pdxA, surA, imp genes from bases 47163 to 57264 (section 5 of 400) of the complete genome	99	115	159
135	1	103	327	emb|X02143|ECPL	Escherichia coli K-12 pldA gene for DR-phospholipase A	97	178	225
135	2	152	409	emb|X02143|ECPL	Escherichia coli K-12 pldA gene for DR-phospholipase A	98	157	258

TABLE 5 (PAI IV) (CONTINUED)

Contig ID	ORF ID	Start (nt)	Stop (nt)	match acession	match gene name	percent ident	HSP nt length	ORF nt length
136	1	122	532	gb|AE000459|	Escherichia coli from bases 4013123 to 4024654 (section 349 of 400) of the complete genome	97	237	411
140	1	576	244	gb|AE000291|	Escherichia coli , asnV, erfK, cobT, cobS, cobU, y152_6, y122_3, y121_3 genes from bases 2060089 to 2072765 (section 181 of 400) of the complete genome	89	329	333
141	1	445	2	gb|AE000291|	Escherichia coli , asnV, erfK, cobT, cobS, cobU, y152_6, y122_3, y121_3 genes from bases 2060089 to 2072765 (section 181 of 400) of the complete genome	77	432	444

TABLE 6 (PAI V)

Putative coding regions of novel *E. coli* PAI V containing known *E. coli* sequences

Contig ID	ORF ID	Start (nt)	Stop (nt)	match accession	match gene name	percent ident	HSP nt length	ORF nt length
3	4	6150	4855	gb|AE000292|	Escherichia coli , yaeA, sbmC, yaeC, sbcB, yaeD, yaeE genes from bases 2072708 to 2083664 (section 182 of 400) of the complete genome	91	129	1296
3	10	8214	7723	emb|X02311|ECIS	E.coli insertion sequence IS3	76	274	492
3	11	7867	8319	emb|Z11606|ECIS	E.coli DNA for insertion sequence IS3	80	378	453
3	12	8462	8157	emb|Z11606|ECIS	E.coli DNA for insertion sequence IS3	90	267	306
3	13	8487	8663	gb|L19084|	Escherichia coli RhsD genetic element; core protein (rhsD) gene, complete cds; complete ORF-D2; complete ORF-D3	96	112	177
4	2	1441	815	gb|AE000498|	Escherichia coli from bases 4493507 to 4503769 (section 388 of 400) of the complete genome	91	577	627
4	3	923	1372	gb|AE000498|	Escherichia coli from bases 4493507 to 4503769 (section 388 of 400) of the complete genome	92	448	450
4	4	2343	1324	gb|AE000498|	Escherichia coli from bases 4493507 to 4503769 (section 388 of 400) of the complete genome	92	244	1020
7	1	3	743	emb|X61239|ECPA	E.coli papABCDEFGHIJK genes for F13 P-pili proteins	100	741	741
7	2	977	615	emb|X61239|ECPA	E.coli papABCDEFGHIJK genes for F13 P-pili proteins	99	363	363
7	3	741	1214	emb|X51704|ECPA	Escherichia coli papJ gene for PapJ protein	98	459	474
8	1	438	4	emb|X60200|ECTN	E. coli transposon Tn1000 (gamma delta) tnpR and tnpA genes for resolvase and transposase	99	435	435
10	1	1932	2426	emb|X61238|ECPR	E.coli prsEFG genes for F13 pili tip proteins	97	462	495
11	1	903	1550	gb|M10133|	E.coli (J96) hlyC, hlyA, hlyB and hlyD genes coding for chromosomal hemolysins C, A, B and D	99	452	648
12	1	2559	1531	gb|U82598|	Escherichia coli genomic sequence of minutes 9 to 12	100	1029	1029
12	2	1594	1860	emb|X13668|ECIS	E.coli insertion element 5 (IS5) DNA	100	267	267
12	3	1858	2235	gb|U95365|	Escherichia coli transposon IS5, transposase (is5B) gene, complete cds	99	354	378
13	1	93	1424	emb|X61239|ECPA	E.coli papABCDEFGHIJK genes for F13 P-pili proteins	99	885	1332
14	9	9832	10515	gb|U09857|	Escherichia coli 4787 o115:v165:f165 fimbrial regulatory f1652I, f1652B and f1652 A genes, complete cds	92	225	684
16	1	1	375	gb|U07174|	Escherichia coli O9:H10:K99 heat resistant agglutinin 1 gene, complete cds	94	320	375
16	2	263	616	gb|U07174|	Escherichia coli O9:H10:K99 heat resistant agglutinin 1 gene, complete cds	98	283	354
17	1	282	4	emb|Y00529|ECPA	E. coli papC gene involved in formation of pap pili	98	240	279
17	2	410	174	emb|Y00529|ECPA	E. coli papC gene involved in formation of pap pili	100	168	237

TABLE 6 (PAI V) (CONTINUED)

Contig ID	ORF ID	Start (nt)	Stop (nt)	match acession	match gene name	percent ident	HSP nt length	ORF nt length
19	1	1	369	gb|AE000418|	Escherichia coli from bases 3550279 to 3561054 (section 308 of 400) of the complete genome	99	347	369
20	10	5401	4829	gb|AE000292|	Escherichia coli , yeeA, sbmC, yeeC, sbcB, yeeD, yeeE genes from bases 2072708 to 2083664 (section 182 of 400) of the complete genome	96	468	573
20	11	4874	5371	gb|AE000292|	Escherichia coli , yeeA, sbmC, yeeC, sbcB, yeeD, yeeE genes from bases 2072708 to 2083664 (section 182 of 400) of the complete genome	96	453	498
20	12	5245	5679	gb|AE000292|	Escherichia coli , yeeA, sbmC, yeeC, sbcB, yeeD, yeeE genes from bases 2072708 to 2083664 (section 182 of 400) of the complete genome	89	235	435
20	13	5732	6139	gb|AE000292|	Escherichia coli , yeeA, sbmC, yeeC, sbcB, yeeD, yeeE genes from bases 2072708 to 2083664 (section 182 of 400) of the complete genome	93	329	408
20	14	6316	5822	gb|AE000292|	Escherichia coli , yeeA, sbmC, yeeC, sbcB, yeeD, yeeE genes from bases 2072708 to 2083664 (section 182 of 400) of the complete genome	95	239	495
20	15	6048	6590	gb|AE000292|	Escherichia coli , yeeA, sbmC, yeeC, sbcB, yeeD, yeeE genes from bases 2072708 to 2083664 (section 182 of 400) of the complete genome	87	406	543
20	16	6569	7075	gb|AE000292|	Escherichia coli , yeeA, sbmC, yeeC, sbcB, yeeD, yeeE genes from bases 2072708 to 2083664 (section 182 of 400) of the complete genome	87	136	507
20	19	8686	9915	gb|M67452|	Escherichia coli lysine decarboxylase (cadB, and cadC, complete cds, and cadA, 5' end) genes	98	1205	1230
20	20	10604	11938	gb|U14003|	Escherichia coli K-12 chromosomal region from 92.8 to 00.1 minutes	98	1308	1335
20	21	11940	12368	gb|M76411|	E.coli cadA gene, 5' cds and cadB and cadC genes, complete cds	100	363	429
21	1	369	4	emb|X03391|ECPA	E. coli major pilu subunit genes genes papI, papB, papA and papH 5'-region	98	201	366
23	1	1	879	gb|U14003|	Escherichia coli K-12 chromosomal region from 92.8 to 00.1 minutes	98	879	879
23	2	900	16	gb|U14003|	Escherichia coli K-12 chromosomal region from 92.8 to 00.1 minutes	98	885	885
23	3	953	1186	emb|X77707|ECCY	E.coli ORF112, DIPZ and ORF191 genes	99	225	234
23	4	1223	2677	emb|X77707|ECCY	E.coli ORF112, DIPZ and ORF191 genes	97	1454	1455
25	1	536	171	emb|X60200|ECTN	E. coli transposon Tn1000 (gamma delta) tnpR and tnpA genes for resolvase and transposase	100	164	366
25	2	1128	562	emb|X60200|ECTN	E. coli transposon Tn1000 (gamma delta) tnpR and tnpA genes for resolvase and transposase	99	459	567
27	1	708	436	emb|X61239|ECPA	E.coli papABCDEFGHIJK genes for F13 P-pili proteins	100	252	273
28	1	309	4	emb|X77707|ECCY	E.coli ORF112, DIPZ and ORF191 genes	98	278	306
28	2	431	213	emb|X77707|ECCY	E.coli ORF112, DIPZ and ORF191 genes	96	150	219

TABLE 6 (PAI V) (CONTINUED)

Contig ID	ORF ID	Start (nt)	Stop (nt)	match acession	match gene name	percent ident	HSP nt length	ORF nt length
30	1	399	4	gb|M26893|	E.coli amidophosphoribosyltransferase (purF) gene, complete cds	98	295	396
31	1	706	170	emb|X56780|ECRR	E.coli terminator sequence of RNA G operon gene	99	513	537
37	1	2	400	gb|M63703|	E.coli pyruvate kinase type II (pykA) gene, complete cds	98	399	399
38	1	463	2	emb|X13463|ECGU	Escherichia coli gutM gene and gutR gene for activator and repressor proteins	99	363	462
42	1	413	3	gb|M64367|	Escherichia coli DNA recombinase (recG) gene, complete cds, spoU gene, 3' end, and gltS gene, 3' end	97	316	411
42	2	115	591	gb|M64367|	Escherichia coli DNA recombinase (recG) gene, complete cds, spoU gene, 3' end, and gltS gene, 3' end	98	266	477
46	1	2	277	emb|X77707|ECCY	E.coli ORF112, DIPZ and ORF191 genes	98	187	276
48	1	1	171	gb|AE000491|	Escherichia coli from bases 4413548 to 4424699 (section 381 of 400) of the complete genome	98	162	171
48	2	105	464	gb|AE000491|	Escherichia coli from bases 4413548 to 4424699 (section 381 of 400) of the complete genome	98	144	360
49	1	2	172	gb|U00800|	Escherichia coli cloning vector Pk184, complete sequence, kanamycin phosphotransferase (kan) and (lacZalpha) genes, complete cds	98	167	171
50	1	414	4	gb|AE000341|	Escherichia coli , glyA, hmpA, glnB, yfhA, yfhG genes from bases 2677406 to 2687636 (section 231 of 400) of the complete genome	99	411	411
52	1	2	307	emb|X60200|ECTN	E. coli transposon Tn1000 (gamma delta) tnpR and tnpA genes for resolvase and transposase	100	284	306
53	1	280	41	gb|M36536|	E.coli htrA gene, complete cds	100	131	240
53	2	558	214	gb|M36536|	E.coli htrA gene, complete cds	99	315	345
54	1	9	263	gb|AE000381|	Escherichia coli from bases 3125914 to 3136425 (section 271 of 400) of the complete genome	94	111	255
55	1	1	675	gb|AE000179|	Escherichia coli , modA, modB, modC, ybhA, ybhE, ybhD genes from bases 794199 to 805132 (section 69 of 400) of the complete genome	98	332	675